

Statistics Workshop Part III

Hidden Markov Models and Significance of Alignment

Silvia Heubach

Department of Mathematics
California State University Los Angeles

Hidden Markov Model (HMM)

In a **Hidden Markov Model**, an underlying Markov chain creates observable output, which is random for each state.

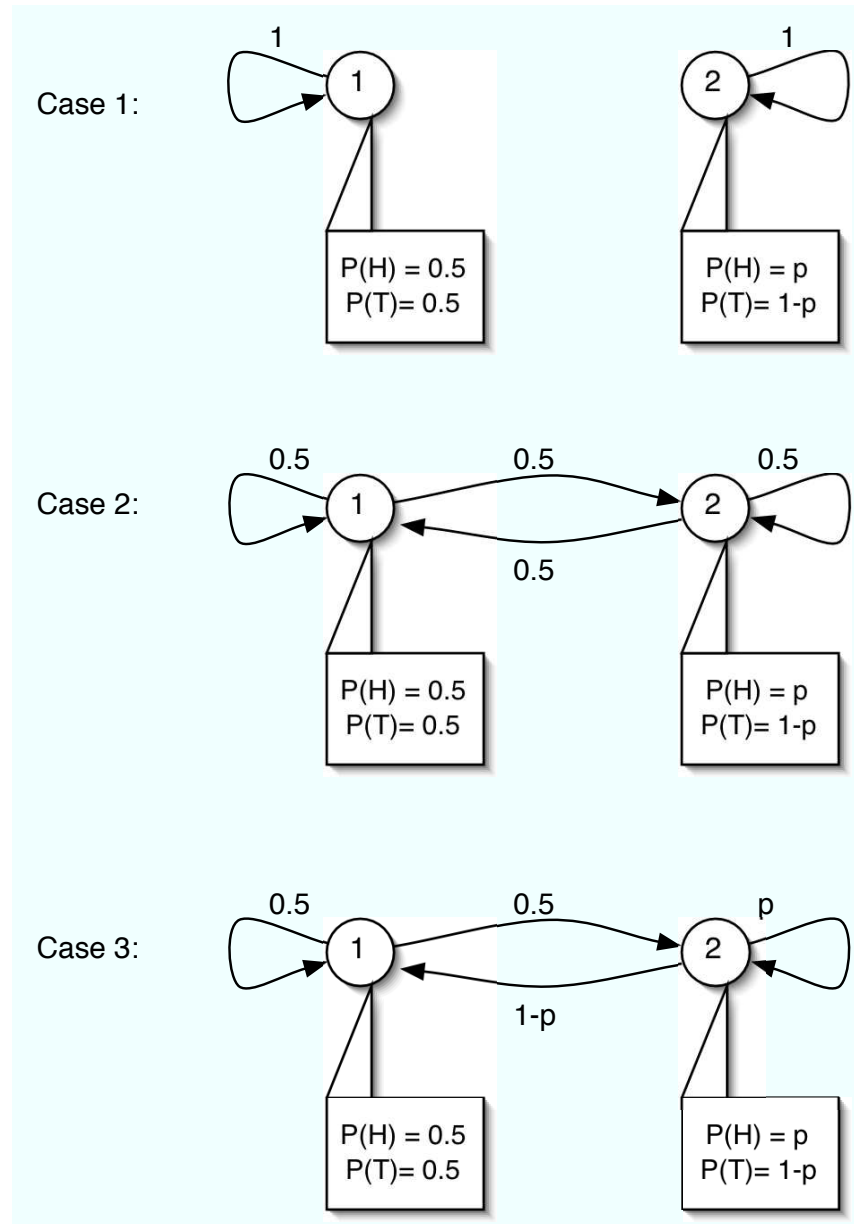
Elements of a HMM

- Finite number of states $\{1, 2, \dots, N\}$
- At each time n , a new state is entered based on transition probabilities that depend only on the previous state (Markov property)
- After each transition, an observable output is produced according to the probability distribution of the current state

Example 1: You are in a room with a barrier. Behind the barrier, a person is performing a coin tossing element. All you are told is the outcomes of the individual tosses HHTHTTHH...

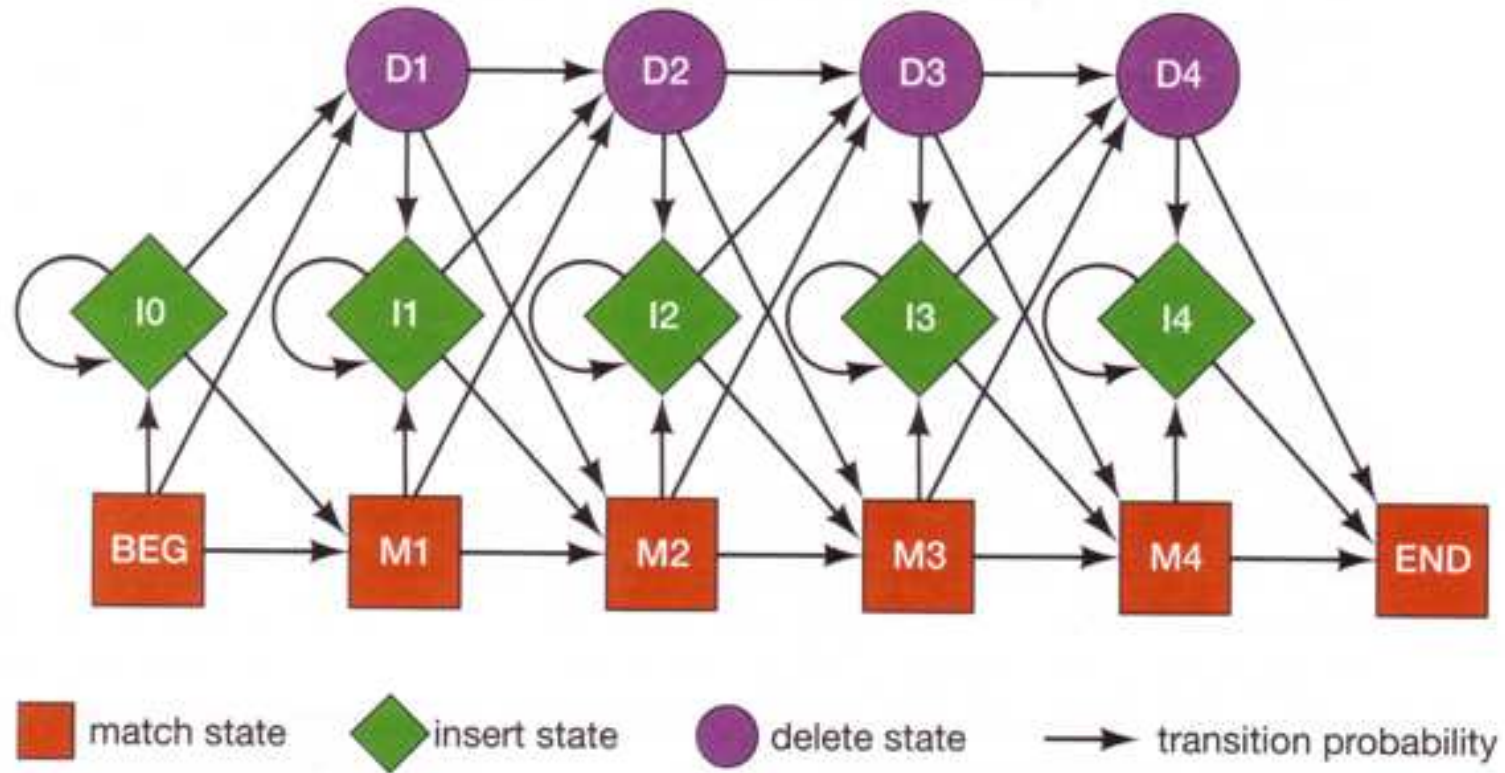
What could be happening behind the barrier: Person chooses one of two coins (one fair, one biased)

- at random and then tosses that coin
- at random for each toss
- and switches to the other coin if the toss results in T



Multiple Sequence Alignment with HMM

- $3n + 3$ states, where n is the average sequence length in the training set
- n match and delete states, $n + 1$ insert states, one begin and one end state
- Training set is used to adjust the transition probabilities until they do no longer change significantly
- From this “trained” model, the most likely path for each sequence to be aligned is computed using the Viterbi algorithm

B. Hidden Markov model for sequence alignment

(From [Mount], Figure 4.16)

- Any sequence can be generated by starting from the BEG state and following a specific path to the END state
- Each match state stores an amino acid distribution (similar to the coin states in the example)
- Each insert state produces a random amino acid for insertion; each delete state produces no output
- Each sequence has an associated probability which is the product of the transition probabilities along the path and the probabilities for obtaining the specific amino acid for the match and insertion states

Sequence Alignment

Example 2:

N	.	F	L	S
N	K	Y	L	T
N	.	F	L	S
Q	.	W	-	T

N K Y L T \leftrightarrow BEG \rightarrow M1 \rightarrow I1 \rightarrow M2 \rightarrow M3 \rightarrow M4 \rightarrow END

Q W T \leftrightarrow BEG \rightarrow M1 \rightarrow M2 \rightarrow D3 \rightarrow M4 \rightarrow END

Associated probabilities

- Transitions from one state to its neighbors are assumed to be equally likely
- Assume match state has uniform distribution for the 20 amino acids $\Rightarrow 0.05$

- Sequence **N K Y L T** with path

BEG \rightarrow M1 \rightarrow I1 \rightarrow M2 \rightarrow M3 \rightarrow M4 \rightarrow END has probability

$$0.33 \times 0.05 \times 0.33 \times 0.05 \times 0.33 \times 0.05 \times 0.33 \times 0.05 \times 0.33 \times 0.05 \times 0.5$$

$$= 6.1 \times 10^{-10}$$

- Resulting values small - use log odds scores

Significance of Alignments

How do we determine whether a sequence alignment is **significant** (biologically meaningful) as opposed to random?

Practical Approach: If the score of the alignment is no better than what might be expected if one used a *random permutation* of the sequence, then it is likely to have arisen by chance.

There are several different measures of significance given by alignment programs:

- *Z*-score
- *P*-value
- *E*-value

We will look at how these values are computed. To do so, we need to compute the distribution of the alignment scores which we obtain by “observing” data from which we can compute mean and standard deviation.

Local alignment of two sequences

- Randomize one of the two sequences many times
- Align each random sequence with the second sequence and score
- Compute distribution of scores

For data base search

- Compare sequence to every sequence in the database and score
- Compute distribution of scores

Local alignment of two sequences

- Randomize one of the two sequences many times
- Align each random sequence with the second sequence and score
- Compute distribution of scores

For data base search

- Compare sequence to every sequence in the database and score
- Compute distribution of scores

Distribution of Scores

A distribution of a random variable can be summarized by single numerical values, such as the mean and the variance of the distribution:

- **mean** μ = weighted sum of the values
- **variance** σ^2 = weighted squared deviation of the values from the mean.

For a discrete random variable,

$$\mu = \sum_x x \cdot p(x) \quad \text{and} \quad \sigma^2 = \sum_x (x - \mu)^2 \cdot p(x).$$

Example 1: X = number of heads in three coin tosses

x	0	1	2	3
$p(x)$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\frac{1}{8}$

$$\mu = \sum_x x \cdot p(x) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = 1.5$$

$$\begin{aligned}\sigma^2 &= \sum_x (x - \mu)^2 \cdot p(x) \\ &= (0 - 1.5)^2 \cdot \frac{1}{8} + (1 - 1.5)^2 \cdot \frac{3}{8} + (2 - 1.5)^2 \cdot \frac{3}{8} + \\ &\quad (3 - 1.5)^2 \cdot \frac{1}{8} = 0.75\end{aligned}$$

Continuous Random Variables

- Random variable X takes values in $[a, b]$
- Distribution described by $[a, b]$ and the **density function** $f(x)$
- Density function is non-negative, but it does not give probabilities. Instead,

$$\begin{aligned} P(X \in [a, b]) &= \text{area under } f(x) \text{ in the interval } [a, b] \\ &= \int_a^b f(x) dx \end{aligned}$$

- Sums are replaced by integrals

$$\mu = \int_a^b x \cdot f(x) \quad \text{and} \quad \sigma^2 = \int_a^b (x - \mu)^2 \cdot f(x).$$

Z-score

Measure of how unusual a match is, as compared to the population. If x is the score of the alignment, then

$$Z\text{-score of } x = \frac{x - \mu}{\sigma}$$

- Computation for the Z -score is simply a standardization of values, such that the random variable Z - score of alignment has mean 0 and standard deviation 1
- Z -score indicates (in terms of standard deviations) how far away the score is from the mean

- $Z\text{-score} = 0 \Rightarrow$ observed similarity is no better than the average in the population, and might have well arisen by chance
- The higher the Z -score, the greater the probability that the observed alignment has not simply arisen by chance
- If scores would be normally distributed (as originally assumed), then significance would occur for $Z\text{-scores} \geq 3$
- Experience suggests that $Z\text{-scores} \geq 5$ are significant \Rightarrow score distribution is NOT a normal distribution

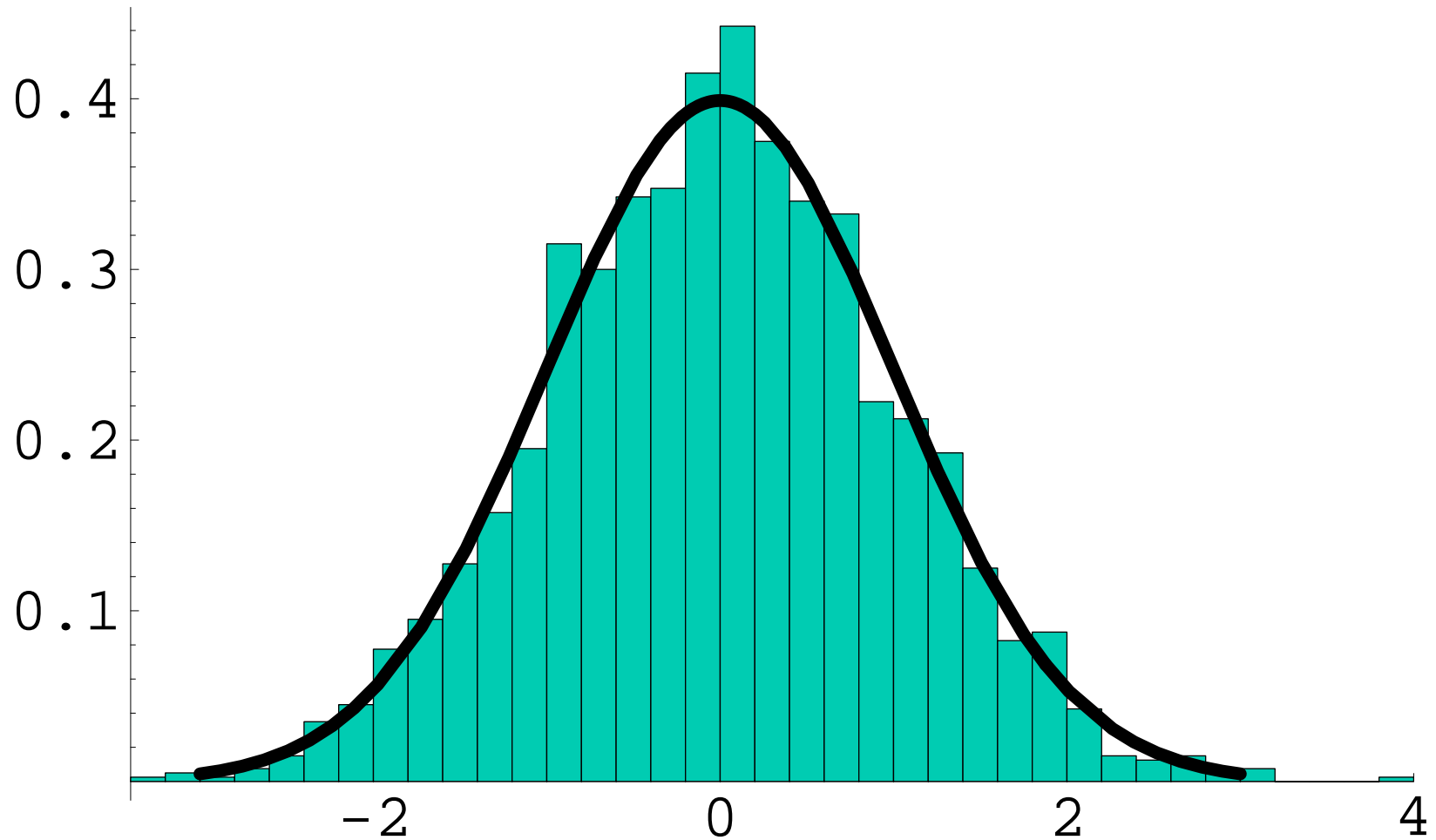
So what type of distribution is it?

...before answering this question, we need to look at how we can get a distribution from data....

Density function from Data

- Collect large amount of data.
- Count how many data points fall into a given interval.
- Scale the height of the bars such that the overall area of the bars totals 1.
- If interval widths are equal, then the area of the bar for a given interval approximates the probability of obtaining data in that interval.

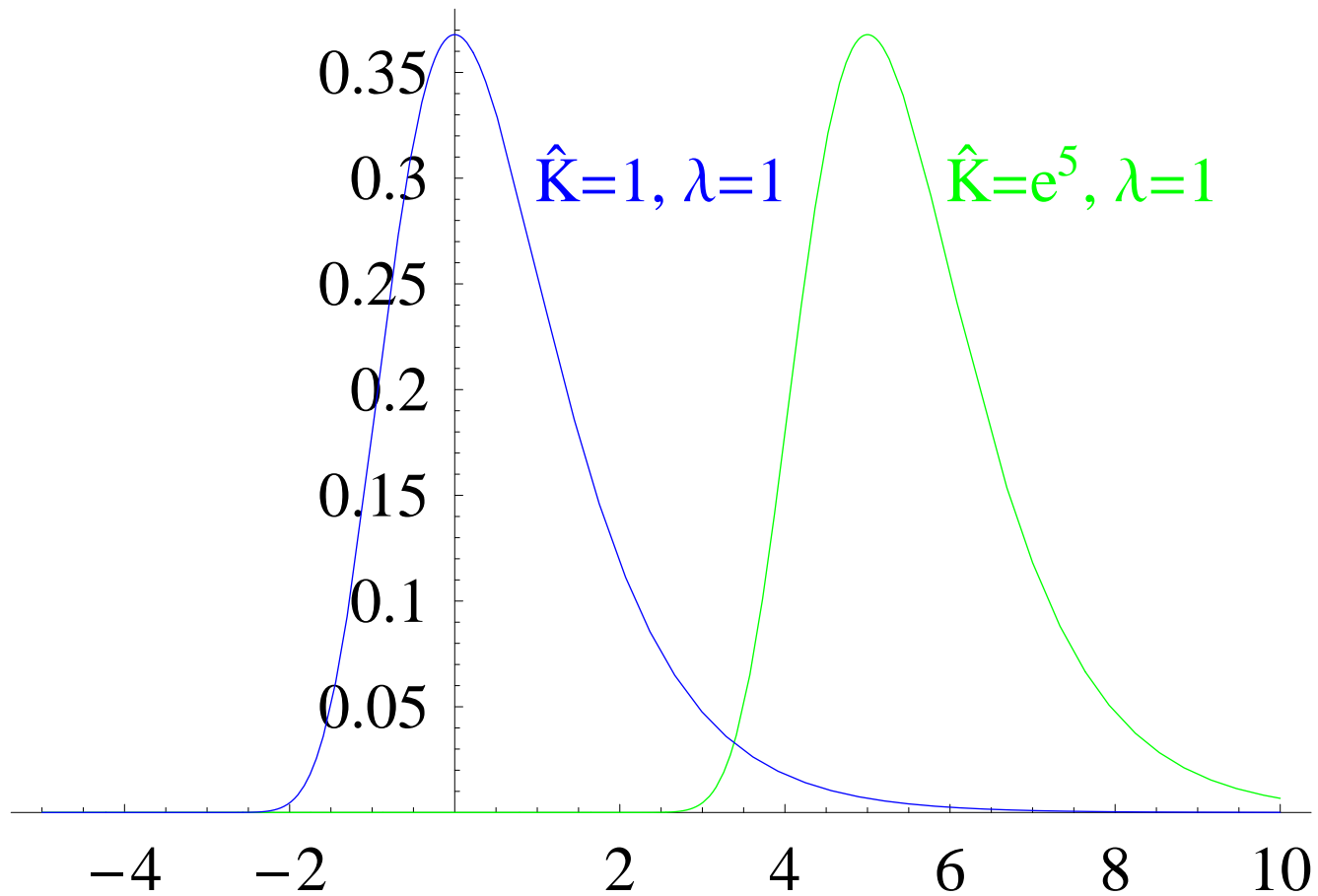
Density function from Data



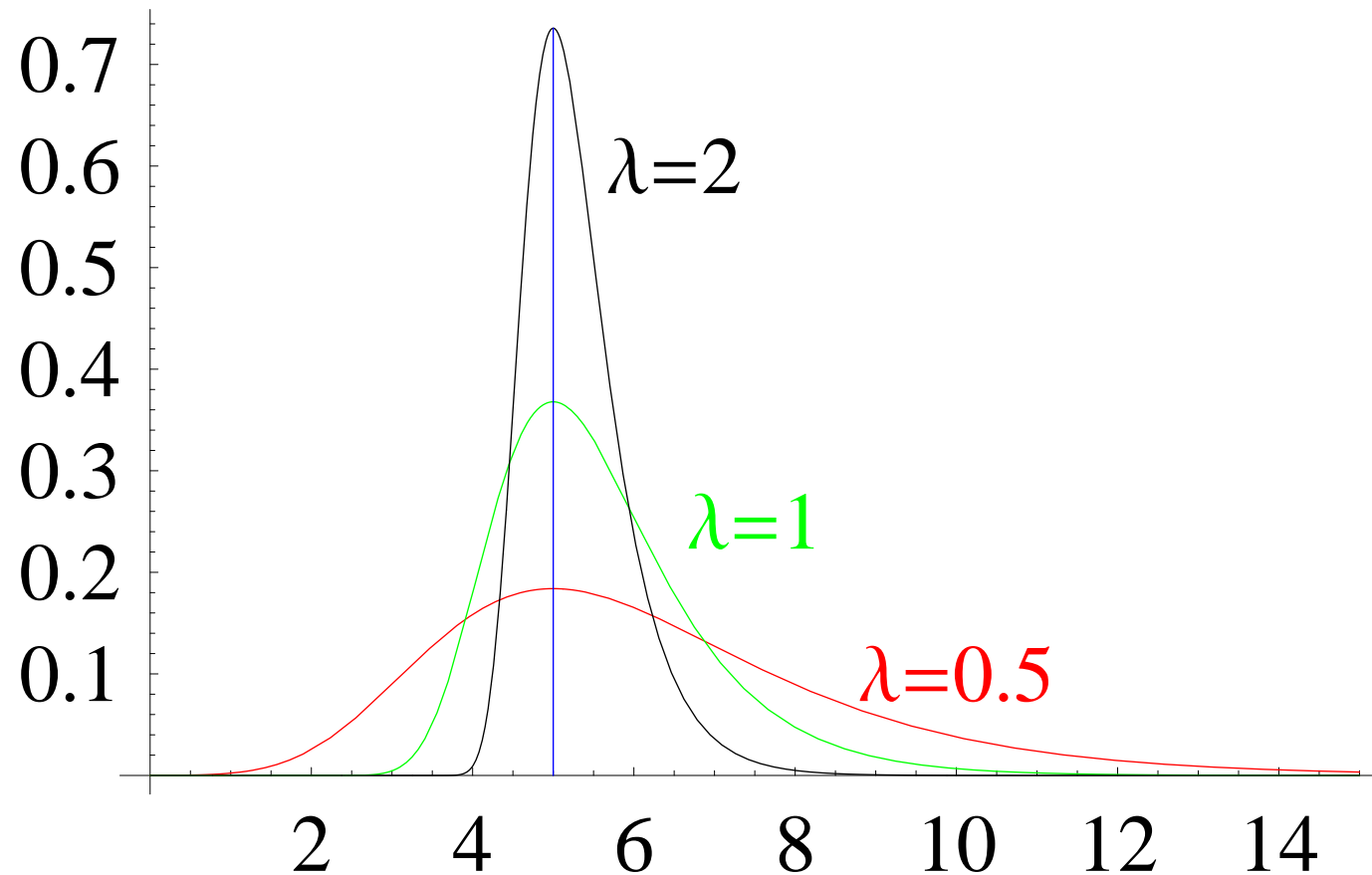
Extreme value distribution

- Density $f(x) = \hat{K} \lambda \exp \left[-\lambda x - \hat{K} e^{-\lambda x} \right]$
- \hat{K} and λ are shape parameters related to maximum and width of distribution
- Maximum occurs at $(\ln \hat{K})/\lambda$
- λ is the **scale** (or **decay**) factor; the larger λ , the more concentrated the values are around the value where the max occurs

EV density with $\lambda=1$ and various values of \hat{K}



EV density with $(\ln \hat{K})/\lambda=5$ and various λ



Extreme Value Distribution

For a score x , the probability that a random score S exceeds x is given by

$$P = P(S \geq x) = 1 - \exp \left[-\hat{K} e^{-\lambda x} \right]$$

This is the ***P-Value*** for the score x . Need to know \hat{K} and λ !

Since the alignment score depends on the scoring matrix, we can obtain two parameters K and λ for each scoring matrix, where $\hat{K} = K \cdot m \cdot n$, and m and n are the lengths of the two sequences that are being aligned

- For some scoring matrices, K and λ have been determined.
- They can also be computed from the “observed” scores as follows: Let s_i be the score of the i th pair, and N be the total number of pairs.

- Sample mean $\bar{x} = \frac{1}{N} \sum_{i=1}^N s_i$

- Sample standard deviation $s^2 = \frac{\sum_{i=1}^N (s_i - \bar{x})^2}{N-1}$

- Then

$$\lambda = \frac{1.2825}{s}$$

$$\hat{K} = \exp[\lambda \bar{x} - .577]$$

Where do these formulas for \hat{K} and λ come from?

$$\hat{K} = \exp[\lambda\bar{x} - .577]$$

- If X has extreme value distribution and m and n are large, then

$$\mu = \int_0^{\infty} x \cdot f(x) dx \approx \frac{0.577 + \ln(\hat{K})}{\lambda}$$

- Estimate μ with \bar{x} and solve for \hat{K} .

Note that the 0.577 is an approximation of Euler's constant γ , which shows up in the evaluation of the integral, independent of the values of \hat{K} or λ .

Example 3:

Suppose two sequences approximately $m = n = 250$ amino acids long are aligned using the PAM250 matrix with high gap penalty. The following alignment is found with alignment score $x = 75$:

FWLEVEGNSMTAPTG
FWLDVQGDSTAPTG

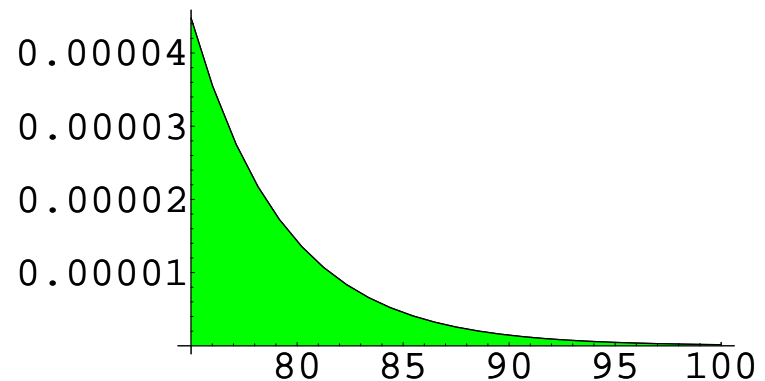
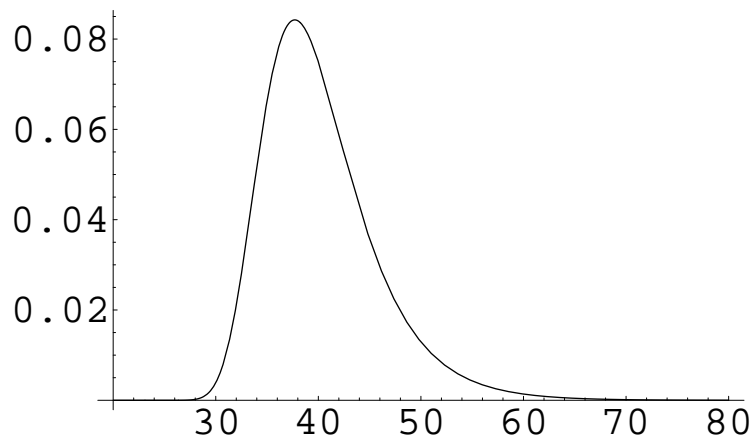
In this case, $K = 0.09$ and $\lambda = 0.229$. Then

$$\begin{aligned} \text{P-value} &= P(S \geq 75) \\ &= 1 - \exp \left[-0.09 \cdot 250 \cdot 250 \cdot e^{-0.229 \cdot 75} \right] \\ &= 0.000195467 \approx 2.0 \times 10^{-4} \end{aligned}$$

Visualization of P -value for above example

The P -value for $x = 75$ is the probability of seeing a score of 75 or more, which is represented by the area underneath the density curve to the right of $x = 75$.

From the graph it is clear that this value is very small, indicating that a score of 75 or higher is very rare, i.e., does not occur by chance and therefore is significant.



Rough guide for interpreting P -values

P -value = P(obtaining score of x or higher).

- $P \leq 10^{-100} \Rightarrow$ exact match
- P in range $10^{-100} - 10^{-50} \Rightarrow$ sequences nearly identical, e.g. alleles or SNPs
- P in range $10^{-50} - 10^{-10} \Rightarrow$ closely related sequences, homology certain
- P in range $10^{-5} - 10^{-1} \Rightarrow$ usually distant relatives
- $P > 10^{-1} \Rightarrow$ match probably insignificant

In our example, P -value $\approx 2.0 \times 10^{-4}$, thus the two sequences are likely to be distant relatives.

P-value versus *E*-value

By the definition of the *P*-value, the smaller the *P*-value, the less likely an outcome has occurred just by chance, i.e., small values of the *P*-value indicate significance. Often, a program will instead report the *E*-value = the expected number of high-scoring segment pairs with score at least x , where

$$E = \hat{K} e^{-\lambda x}.$$

Note that

$$P\text{-value} = 1 - \exp[-E].$$

In our example, $E = 0.000195486$, which is almost identical to the *P*-value.

Rough guide for interpreting E -values

E -value = expected number of sequences that give the same x -score or better when aligned with a random sequence.

Rough guide for interpreting E -values:

- $E \leq 0.02 \Rightarrow$ sequences probably homologous
- E between 0.02 and 1 \Rightarrow homology cannot be ruled out
- $E > 1$ you would expect this good a match just by chance

In our example, the E -value was $E = 0.000195486$, i.e., the two sequences are likely to be homologous.