Some probability and stats refreshers

**Iterated expectation:** $E(g(X,Y)) = E(E(g(X,Y)|X))$.

**Total Variation:** $\text{Var}(Y) = E(\text{Var}(Y|X)) + \text{Var}(E(Y|X))$.

Example: $Y=\text{height}$, $X=\text{sex}$, $Y|X=F \sim N(165,25)$, $Y|X=M \sim N(175,35)$, $\mathbb{P}(X=F) = 0.5$.

Then: $E(Y) = E(E(Y|X)) = 0.5 \times 165 + 0.5 \times 175 = 170$.

$\text{Var}(Y) = 0.5 \times 25 + 0.5 \times 35 + 0.5 \times 5^2 + 0.5 \times 5^2 = 55$.

**Exponential distribution:** $X \sim \text{exp}(\lambda)$, $\mathbb{P}(X \geq t) = e^{-\lambda t}$.

It has the lack of memory property:

$$\mathbb{P}(X > u + a|X > a) = \frac{e^{-\lambda (u + a)}}{e^{-\lambda a}} = e^{-\lambda u},$$

so regardless of our waiting time so far $a$, the probability we will wait $u$ longer is fixed.

Note that other distributions don’t have this property, for example for $X \sim N(0,1)$:

$$\mathbb{P}(X \geq 0|X \geq -5) \approx 0.5, \quad \mathbb{P}(X \geq 5|X \geq 0) = 2\mathbb{P}(X \geq 5) \approx 0.$$

**Renewal process:** Assume $X_i \sim \text{exp}(\lambda)$ i.i.d and we wait $X_1$ for the first event, then $X_2$ for the second etc. In a given time $T$ the number of events:

$$k(T) = \max \{i : X_1 + \ldots + X_i \leq T\}.$$  

Claim: $k(T) \sim \text{Poiss}(\lambda T)$.

Partial proof:

$$\mathbb{P}(k = 0) = \mathbb{P}(X_1 > T) = e^{-\lambda T}.$$  

$$\mathbb{P}(k = 1) = \int_0^T f(X_1 = t)\mathbb{P}(X_2 > T - t)dt = \int_0^T \lambda e^{-\lambda t}e^{-\lambda (T-t)}dt = e^{-\lambda T} \int_0^T \lambda dt = \lambda T e^{-\lambda T}.$$
Poisson and binomial: If \( n \) is big, \( p \) is small then \( \text{Bin}(n, p) \approx \text{Pois}(np) \).

Intuition: \( n \) independent increments, in each one fixed probability of event \( \Rightarrow \) a memoryless process, approximately in continuous time.

Example: Molecular clock calculations

Assume now we have \( n \) generations of mutations father\( \rightarrow \)son\( \rightarrow \)grandson etc.
Assume every generation has fixed probability \( p \) of mutation ("Molecular clock").
Then number of mutations \( k \) in \( n \) generations: \( \text{Bin}(n, p) \approx \text{Pois}(np) \).

Rather than in discrete generations, we can also think of this continuously, where a mutation can happen in every point in time at a fixed rate \( \lambda \), so the waiting time for mutation has \( \exp(\lambda) \) distribution with mean \( 1/\lambda \).
If we now assume generation length is \( t_0 \), then the number of mutations in \( n \) generations has a \( \text{Pois}(nt_0\lambda) \) distribution, that is the binomial \( p \) above is \( t_0\lambda \).

When we look at genetic sequences and observe differences the classical problems are:

1. **Calibration:** Given time (\( n \) or \( nt_0 \)) estimate the mutation rate \( \lambda \) or \( p \).

2. **Time estimation:** Given the rate \( \lambda \) estimate the time \( T = nt_0 \) separating between sequences of species.