Module 3B: Thinking in Distributions

Building block for Hypothesis Testing

Agenda:

- Major distributions:
  - **Discrete Distributions**
    - Bernoulli
    - Binomial
    - Poisson
    - Hypergeometric
  - **Continuous Distributions**
    - Normal
    - Uniform
    - Exponential
    - Gamma
- Interactive simulations
- **Central Limit Theorem**
  - **Sampling Distribution of the mean**

Poisson Distribution

The *Poisson distribution* is a discrete distribution modeling the number of times an event occurs in a time interval, given that the average number of events occurring in the interval is known. You can think of the Poisson distribution as a special case of the Binomial distribution but with a **really large number of intervals and a really small probability of success in any given one interval** (in more math-y speak: n approaches infinity and p approaches 0). You only need to know the mean number of an event – it is useful to not need to know the exact number of intervals or where the events happened to describe, for instance, how mutations are distributed along genealogies. There is an accessible explanation of how to derive the Poisson distribution from the binomial distribution here (warning: there are limits involved): https://medium.com/@andrew.chamberlain/deriving-the-poisson-distribution-from-the-binomial-distribution-840cc1668239

$$P(x;\mu) = \frac{e^{-\mu}\mu^{x}}{x!}$$

Mean = Variance = μ

Question: What is the probability of having 110 novel mutations (mutations that neither of your parents have) if the mean mutation rate of the human genome is 115?

https://probstats.org/poisson.html

Hypergeometric Distribution

- This is used in tag-and-release programs.
- Basis for one-tailed version of Fisher's exact test (we will see this later)

https://probstats.org/hypgeom.html

$$P(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}}$$

Question: A research colony has 50 mice, of which 12 carry a CRISPR knockout of Gene X and the remaining 38 are wild-type. You randomly select 8 mice for a behavioral assay without replacement. What is the probability that exactly 3 of the selected mice carry the Gene X knockout?

https://en.wikipedia.org/wiki/Hypergeometric_distribution

Uniform Distribution

https://probstats.org/uniform_discrete.html

- used as an “information-less” prior in Bayes’ Formula

Formula: $f(x) = \frac{1}{max-min}$

Mean: $\frac{max+min}{2}$

Variance: $\frac{(max+min)^2}{12}$

Question: A chromosome is 2 Morgans in length (recall that one Morgan corresponds to the genomic distance of a mean of one crossover event). Is a mean observed position of the crossover is at 1 Morgan with a variance of 1/2 consistent with a uniform distribution of crossover events along the 2 Morgan chromosome?

Normal Distribution

- Bedrock of inferential statistics
- Approximate the Binomial Distribution (n >30).
- Phenomenon that result from many additive small effects processes are normally distributed --- this is very common in biological processes (Single mutation Mendelian traits are the exception rather than the rule)
- **Central limit theorem:** means of samples of random variables from other distributions (not normal distributions) can approach a normal distribution as then sample size increases.

  Formula: $f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}$

  Mean: μ

  Variance: σ^2

  (The complicated denominator ensures that the distribution integrates to 1.)

- https://probstats.org/normal.html

Exponential Distribution

https://probstats.org/exponential.html

Formula: $\boldsymbol{f(x) = \alpha e^{-\alpha x}}$

Mean: $\frac{\mathbf{1}}{\boldsymbol{\alpha}}$

Variance: $\frac{\mathbf{1}}{\boldsymbol{\alpha^2}}$

Question: A bee is foraging and stops at flowers at a constant rate, $\alpha = 0.05$ per meter. What is the mean distance travelled between flowers?