Review

- Random variables
  - Discrete: Probability mass function (pmf)
  - Continuous: Probability density function (pdf)

- Probability distributions that are popular in genetics and genomics
  - Binomial distribution
  - Hypergeometric distribution

- Why is it important to learn about probability distributions?
  - Given the pdf or pmf of a rv. \( X \),
  - We can compute the probability of various events, mean/variance of \( X \), without having to perform experiments & count from the data
  - We can simulate the real system and get the data

Today...

- More on discrete distributions
  - Poisson distribution

- Continuous distributions
  - Uniform distribution
  - Exponential distribution
  - Gamma distribution
  - Normal distribution

- R session
  - Working with distributions in R

Poisson Distribution

- Probability of a given number of events \( (X = i) \) occurring in a fixed interval of time and/or space \( (t) \)

- Assumption: Events occur with a known average rate \( (\lambda) \) and independently of the time since the last event

- Any simple example?
Intuitive Example

- Say that you typically get 10 e-mails per day.
- That becomes the expectation, but there will be a certain spread: sometimes a little more, sometimes less.
- Given only the average rate for a certain period of observation (e.g. # e-mails per day),
- Poisson distribution specifies how likely it is that the count will be 3, 5 or 11, or any other number, during one period of observation (e.g. 1 week).
- It predicts the degree of spread around a known average rate of occurrence.

Poisson Distribution

- A rv $X$ follows a Poisson distribution if the pmf of $X$ is:

$$P(X = i) = e^{-\lambda} \frac{\lambda^i}{i!}$$

For $i = 0, 1, 2, 3, \ldots$

- e.g. $X = \#$ e-mails per week
- Given $\lambda$ (a rate per 1 unit time), $\lambda = \frac{ext}{t}$ expected number of events per unit time $t$
- e.g. Given 4 e-mails per day, how many e-mails per week?
- $E(X) = \text{Var}(X) = \lambda$

Poisson RV: Example 1

- The number of crossovers, $X$, between two markers is $X \sim \text{Poisson}(\lambda=d)$

$$P(X = i) = e^{-d} \frac{d^i}{i!}$$

$P(X = 0) = e^{-d}$

$P(X \geq 1) = 1 - e^{-d}$

Poisson RV: Example 2

- Recent work in Drosophila suggests the spontaneous rate of deleterious mutations is $\sim 1.2$ per diploid genome.

Thus, let’s tentatively assume $X \sim \text{Poisson}(\lambda = 1.2)$ for humans. What is the probability that an individual has 12 or more spontaneous deleterious mutations?

$$P(X \geq 12) = 1 - \sum_{i=0}^{11} e^{-1.2} \frac{1.2^i}{i!}$$

$$= 6.17 \times 10^{-9}$$
Poisson RV: Example 3

- Suppose that a rare disease has an incidence of 1 in 1000 people per year. Assuming that members of the population are affected independently.

- Find the probability of $k$ cases in a population of 10,000 (followed over 1 year) for $k=0,1,2$.

- The expected value (mean) $\lambda = 0.001 \times 10,000 = 10$

\[
P(X = 0) = \frac{(10)^0 e^{-10}}{0!} = .0000454
\]
\[
P(X = 1) = \frac{(10)^1 e^{-10}}{1!} = .000454
\]
\[
P(X = 2) = \frac{(10)^2 e^{-10}}{2!} = .00227
\]

Poisson Distribution

- Useful in studying rare events
  - $\alpha$ is very low; $t$ is very large
  - $\lambda (= \alpha t)$ is of intermediate magnitude

- Poisson distribution approximates the binomial distribution when $n$ (# trials) is large and $p$ (change of success) is small
  - Safely approximates a binomial experiment when $n > 100$, $p < 0.01$, $np = \lambda < 20$

Poisson vs. Binomial Distribution

- Given $n$ trials and success rate of $p$, what is the probability that there are $k$ successes?
  - Binomial distribution:
    
    \[
P[X = k] = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}
    \]

  - In such cases $n$ is very large and $p$ is very small, the distribution may be approximated by the less cumbersome Poisson distribution with $\lambda (= np)$
    
    \[
P[X = i] = e^{-\lambda} \frac{\lambda^i}{i!} e^{-\frac{(np)^i}{\lambda^i}}
    \]

  - This is sometimes known as the law of rare events

Poisson vs. Binomial Distribution

- All distribution have a mean of 5
Proof (not to be covered in class)

- Poisson distribution is a limiting case of binomial distribution
- For $n$ trials of events with probability $p$ and $\lambda = np$,
  $$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n(n-1)(n-2)...(3)(2)(1)}{(n-k)(n-k-1)...(2)(1)} \left( \frac{\lambda}{n} \right)^k \left( 1 - \frac{\lambda}{n} \right)^{n-k}$$
- Since $\lim_{n \to \infty} \left( 1 - \frac{\lambda}{n} \right)^{n-k} = \exp \left( -\frac{\lambda}{n} \right)$, we have
  $$\left( 1 - \frac{\lambda}{n} \right)^{n-k} = \left( 1 - \frac{\lambda}{n} \right)^n \to e^{-\lambda}$$
  $$P(X = k) = \frac{n(n-1)(n-2)...(n-k+1)}{n^k} \cdot \frac{1}{k!} e^{-\lambda}$$
- and as $n$ gets large the fraction in front goes to $1/k!$ so that
  $$\frac{1}{k!} e^{-\lambda}$$

Poisson or Binomial Distribution?

- If a mean or average probability of an event happening per unit time/space is given and you’re asked to calculate a probability of $k$ events happening in a given time/space, then ...
- If, on the other hand, an exact probability of an event happening is given, or implied, and you are asked to calculate the probability of this event happening $k$ times out of $n$, then ...

Quiz #1: Poisson or Binomial?

- A typist makes on average 2 mistakes per page. What is the probability of a particular page having no errors on it?

Quiz #2

- A computer crashes once every 2 days on average. What is the probability of there being 2 crashes in one week?
Quiz #3
- Components are packed in boxes of 20. The probability of a component being defective is 0.1. What is the probability of a box containing 2 defective components?

Uniform Distribution
- Suppose we have a telephone line and we listen to it for one hour. A call comes in at a random time in that hour. How can we make a histogram of when the call comes in?
  - If we divide 1 hour into four 15-minute periods, the histogram of probabilities of getting a call is:

Uniform Distribution
- If we divide it into 60 1-minute blocks, we have:

Continuous Distribution
- The solution is to not record probabilities of being “at” a value, but a density function which shows the relative probabilities of being in different regions, scaled so that the area under it is 1.
  - Then we can use it to compute the probability of being in any interval, by integrating the function between those bounds.
Continuous Distribution

- More generally, here is the uniform distribution (or uniform density) between $a$ and $b$:

\[
\begin{align*}
1/(b-a) \\
\hline
-a & b
\end{align*}
\]

Exponential Distribution

- A r.v $X$ follows an exponential distribution if the pdf of $X$ is:

\[
P(x) = \begin{cases} 
\lambda e^{-\lambda x}, & x \geq 0 \\
0, & x < 0 
\end{cases}
\]

- $E(X) = 1/\lambda$
- $Var(X) = 1/\lambda^2$

- The cumulative distribution is given by:

\[
F(x) = \begin{cases} 
1 - e^{-\lambda x}, & x \geq 0 \\
0, & x < 0 
\end{cases}
\]

More Continuous Distributions

- Uniform distribution
- Exponential distribution
- Gamma distribution
- Normal distribution

Relationship To Poisson Distribution

- Poisson process: events occur continuously at a constant average rate ($\lambda$) – the average number of arrivals per unit time,

\[P\{X = i\} = e^{-\lambda} \frac{\lambda^i}{i!}\]

- The exponential distribution describes the time between events in a Poisson process
Gamma Distribution

- Waiting time to the $k^{th}$ telephone call
- Related to the Poisson distribution
  - If we wait a fixed amount of time, each little chunk of time is like the toss of a coin with a very small probability of Heads, and we receive a Poisson number of telephone calls.
  - If instead we wait until the $k^{th}$ call comes, the waiting time is Gamma-distributed.

Normal Distribution

- “Most important” probability distribution
- Many rv’s are approximately normally distributed
- Even when they aren’t, their sums and averages often are Central Limit Theorem (CLT)

Gamma Distribution

- The Gamma distribution has 2 parameters
  - Shape parameter $k$: Number of the calls you are waiting
  - Scale parameter $\theta$: Rate of phone calls expected
  - The parameters are continuous; so also you can get a Gamma density for fractional phone calls too

- The pdf of Gamma distribution is

$$P(x) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta}$$

Normal Distribution

- pdf of normal distribution:
  $$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

  - standard normal distribution ($\mu = 0, \sigma^2 = 1$):
  $$f(z; 0, 1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

  - cdf of $Z$
  $$P(Z \leq z) = \int_{-\infty}^{z} f(y; 0, 1) \, dy$$
Standardizing Normal RV

- If $X$ has a normal distribution with mean $\mu$ and standard deviation $\sigma$, we can standardize to a standard normal rv:

$$Z = \frac{X - \mu}{\sigma}$$

$$f(z;0,1) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

1 Digress: Sample Distributions

- Before data is collected, we regard observations as random variables $X_1, X_2, ..., X_n$.
- This implies that until data is collected, any function (statistic) of the observations (mean, sd, etc) is also a random variable.
- Thus, any statistic, because it is a random variable, has a probability distribution – referred to as a sample distribution.
- Let’s focus on the sampling distribution of the mean, $\bar{X}$.

Behold The Power of the CLT

- Let $X_1, X_2, ..., X_n$ be an iid random sample from a distribution with mean $\mu$ and standard deviation $\sigma$. If $n$ is sufficiently large:

$$\bar{X} \sim \text{N}(\mu, \frac{\sigma}{\sqrt{n}})$$

Example

- If the mean and standard deviation (sd) of serum iron values from healthy men are 120 and 15 mgs per 100ml, respectively.
- What is the probability that a random sample of 50 normal men will yield a mean between 115 and 125 mgs per 100ml?

First, calculate mean and sd to normalize (120 and 15/sqrt(50))

$$p(115 \leq \bar{X} \leq 125) = p\left(\frac{115 - 120}{15/\sqrt{50}} \leq \frac{\bar{X} - 120}{15/\sqrt{50}} \leq \frac{125 - 120}{15/\sqrt{50}}\right)$$

$$= p(-2.36 \leq z \leq 2.36)$$

$$= p(z \leq 2.36) - p(z \leq -2.36)$$

$$= 0.9909 - 0.0091$$

$$= 0.9818$$