Stat 200A: Homework 1

1 Work out the rare disease example.

$$\begin{array}{lll} P(D) = 1\% & P(+|D) = 90\% & P(-|D) = 10\% \\ P(N) = 99\% & P(+|N) = 10\% & P(-|N) = 90\% \end{array}$$

The probability that one has such kind of disease given he gets '+' in the medical test is

$$\begin{aligned} P(D|+) &= \frac{P(D \cap +)}{P(+)} \\ &= \frac{P(D \cap +)}{P(D \cap +) + P(N \cap +)} \\ &= \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|N)P(N)} \\ &= \frac{0.9 \times 0.01}{0.9 \times 0.01 + 0.1 \times 0.99} \\ &= \frac{1}{12} \end{aligned}$$

and similarly we have

$$\begin{aligned} P(D|-) &= \frac{P(-|D)P(D)}{P(-|D)P(D) + P(-|N)P(N)} = \frac{1}{892} \\ P(N|+) &= \frac{P(+|N)P(N)}{P(+|N)P(N) + P(+|D)P(D)} = \frac{11}{12} \\ P(N|-) &= \frac{P(-|N)P(N)}{P(-|N)P(N) + P(-|D)P(D)} = \frac{891}{892} \end{aligned}$$

The results tell that because the prior probability that one has such kind of disease $P(D)$ is very small, even one gets '+' in the medical test, the chance that he/she has the disease is still small; on the contrary, even one gets '$-$' in the medical test, the chance that he/she doesn't have the disease is still large.

2 Work out the smoking habit example.

Smoking habit (cigarette or pipe) and health status (healty or unhealthy) are conditional independent given the age of the smokers.
For young smokers (Y):

Pr.	healthy (H)	unhealthy (U)
cigarette (C)	0.81	0.09
pipe (P)	0.09	0.01

For old smokers (O):

Pr.	healthy (H)	unhealthy (U)
cigarette (C)	0.01	0.09
pipe (P)	0.09	0.81

Suppose $P(Y) = P(O) = \frac{1}{2}$, then

$$\begin{aligned}
P(H|C) &= \frac{P(H\cap C)}{P(H\cap C) + P(U\cap C)} \\
&= \frac{P(H\cap C\cap Y) + P(H\cap C\cap O)}{P(H\cap C\cap Y) + P(H\cap C\cap O) + P(U\cap C\cap Y) + P(U\cap C\cap O)} \\
&= \frac{P(H\cap C|Y)P(Y) + P(H\cap C|O)P(O)}{P(H\cap C|Y)P(Y) + P(H\cap C|O)P(O) + P(U\cap C|Y)P(Y) + P(U\cap C|O)P(O)} \\
&= \frac{0.81\times 0.5 + 0.01\times 0.5}{0.81\times 0.5 + 0.01\times 0.5 + 0.09\times 0.5 + 0.09\times 0.5} = 0.82 \\
P(H|P) &= \frac{P(H\cap P|Y)P(Y) + P(H\cap P|O)P(O)}{P(H\cap P|Y)P(Y) + P(H\cap P|O)P(O) + P(U\cap P|Y)P(Y) + P(U\cap P|O)P(O)} \\
&= \frac{0.09\times 0.5 + 0.09\times 0.5}{0.09\times 0.5 + 0.09\times 0.5 + 0.01\times 0.5 + 0.81\times 0.5} = 0.18
\end{aligned}$$

$P(H|C) > P(H|P)$ but it doesn't tell that smoking habit has effect on health status, since age has influence on both, and

$$\begin{aligned}
P(H|C\cap Y) &= \frac{P(H\cap C\cap Y)}{P(H\cap C\cap Y) + P(U\cap C\cap Y)} \\
&= \frac{P(H\cap C|Y)P(Y)}{P(H\cap C|Y)P(Y) + P(U\cap C|Y)P(Y)} \\
&= \frac{P(H\cap C|Y)P(Y)}{P(H\cap C|Y)P(Y) + P(U\cap C|Y)P(Y)} \\
&= \frac{0.81\times 0.5}{0.81\times 0.5 + 0.09\times 0.5} = 0.9 \\
P(H|P\cap Y) &= \frac{P(H\cap P|Y)P(Y)}{P(H\cap P|Y)P(Y) + P(U\cap P|Y)P(Y)} \\
&= \frac{0.09\times 0.5}{0.09\times 0.5 + 0.01\times 0.5} = 0.9
\end{aligned}$$

where $P(H|C\cap Y) = P(H|P\cap Y)$ and similarly we have $P(H|C\cap O) = P(H|P\cap O) = 0.1$ for the old smokers group.

3 Derive the exponential distribution and Poisson distribution from Poisson process.

Suppose we have a Poisson process with rate λ, then the probability that the first event arrives in the kth Δt is

$$P(k) = (1 - \lambda\Delta t)^{k-1}\lambda\Delta t$$

thus the distribution of the waiting time t until the first arrival is

$$\begin{aligned}
f(t) &= \lim_{\Delta t\to 0} \frac{P(k)}{\Delta t} \\
&= \lim_{\Delta t\to 0} \frac{(1-\lambda\Delta t)^{k-1}\lambda\Delta t}{\Delta t} \\
&= \lambda \lim_{\Delta t\to 0} \frac{(1-\lambda\Delta t)^{\frac{t}{\Delta t}}}{1-\lambda\Delta t} \\
&= \lambda e^{-\lambda\Delta t\frac{t}{\Delta t}} \\
&= \lambda e^{-\lambda t}
\end{aligned}$$

This is the exponential distribution. Then the probability that k events arrive during n intervals of Δt is

$$P(k) = \binom{n}{k}(\lambda\Delta t)^k(1-\lambda\Delta t)^{n-k}$$

thus the distribution of the number of arrivals during an interval of length t is

$$\begin{aligned}
f(k) &= \lim_{\Delta t\to 0} P(k) \\
&= \lim_{\Delta t\to 0} \binom{n}{k}(\lambda\Delta t)^k(1-\lambda\Delta t)^{n-k} \\
&= \lim_{\Delta t\to 0} \frac{\prod\limits_{i=0}^{k-1}(\frac{t}{\Delta t}-i)}{k!}(\lambda\Delta t)^k(1-\lambda\Delta t)^{\frac{t}{\Delta t}-k} \\
&= \frac{\lambda^k}{k!}\lim_{\Delta t\to 0}\prod_{i=0}^{k-1}(t-i\Delta t)\frac{(1-\lambda\Delta t)^{\frac{t}{\Delta t}}}{(1-\lambda\Delta t)^k} \\
&= \frac{\lambda^k}{k!}t^k e^{-\lambda\Delta t\frac{t}{\Delta t}} \\
&= \frac{(\lambda t)^k}{k!}e^{-\lambda t}
\end{aligned}$$

This is Poisson distribution.

4 Derive the normal distribution from random walk. Please check out quincunx and Einstein's random walk.

Suppose we have a random walk with each step $Z_i \in \{+\Delta x, -\Delta x\}$, where $P(Z_i = +\Delta x) = P(Z_i = -\Delta x) = \frac{1}{2}$. Starting from zero, the probability that we are at $k\Delta x$ after n steps equals the probability of taking n_R steps towards right and n_L steps towards left, such that $n_R + n_L = n$ and $n_R - n_L = k$, so it should be

$$P(k) = \frac{\binom{n}{n_R}}{2^n} = \frac{\binom{n}{\frac{n+k}{2}}}{2^n} = \begin{cases} \frac{n!}{2^n(\frac{n-k}{2})!(\frac{n+k}{2})!} & n+k \text{ is even} \\ 0 & n+k \text{ is odd} \end{cases}$$

We also have the diffusion equation $\Delta x = \sigma\sqrt{\Delta t} = \sigma\sqrt{\frac{t}{n}}$, where t is the total time for the n steps. So fixing t, when $\Delta x \to 0$, $n \to \infty$. Thus the distribution of the state of the random walk at time t is

$$\begin{aligned}
f(x) &= \lim_{\Delta x \to 0} \frac{P(k)}{\Delta x} \\
&= \lim_{\Delta x \to 0} \frac{n!}{2^n (\frac{n-k}{2})! (\frac{n+k}{2})!} \cdot \frac{1}{2} \cdot \frac{1}{\Delta x} \\
&= \lim_{\Delta x \to 0} \frac{\sqrt{2\pi} n^{n+\frac{1}{2}} e^{-n}}{2^n \sqrt{2\pi} \left(\frac{n-k}{2}\right)^{\frac{n-k}{2}+\frac{1}{2}} e^{-\frac{n-k}{2}} \sqrt{2\pi} \left(\frac{n+k}{2}\right)^{\frac{n+k}{2}+\frac{1}{2}} e^{-\frac{n+k}{2}} \cdot 2\Delta x} && \text{(Stirling Formula)} \\
&= \lim_{\Delta x \to 0} \frac{n^{n+\frac{1}{2}} e^{-n}}{2^n \sqrt{2\pi} \cdot \frac{(n-k)^{\frac{n-k}{2}+\frac{1}{2}} (n+k)^{\frac{n+k}{2}+\frac{1}{2}}}{2^{n+1}} \cdot e^{-\frac{n-k}{2} - \frac{n+k}{2}} \cdot 2\Delta x} \\
&= \lim_{\Delta x \to 0} \frac{n^{n+\frac{1}{2}}}{\sqrt{2\pi} \sqrt{n^2-k^2} (n-k)^{\frac{n-k}{2}} (n+k)^{\frac{n+k}{2}} \sigma \sqrt{\frac{t}{n}}} \\
&= \lim_{\Delta x \to 0} \frac{1}{\sqrt{2\pi\sigma^2 t} (1-\frac{k}{n})^{\frac{n-k}{2}} (1+\frac{k}{n})^{\frac{n+k}{2}}} \\
&= \frac{1}{\sqrt{2\pi\sigma^2 t}} e^{-(-\frac{k}{n} \cdot \frac{n-k}{2} + \frac{k}{n} \cdot \frac{n+k}{2})} \\
&= \frac{1}{\sqrt{2\pi\sigma^2 t}} e^{-\frac{k^2}{n}} \\
&= \frac{1}{\sqrt{2\pi\sigma^2 t}} e^{-\frac{\left(\frac{x}{\Delta x}\right)^2}{n}} \\
&= \frac{1}{\sqrt{2\pi\sigma^2 t}} e^{-\frac{\frac{x^2 n}{\sigma^2 t}}{n}} \\
&= \frac{1}{\sqrt{2\pi\sigma^2 t}} e^{-\frac{x^2}{\sigma^2 t}} = \mathcal{N}(0, \sigma^2 t)
\end{aligned}$$