Probability Theory Lecture Notebook

Contents

Basic useful formula

Special distribution

Binomial disributions
A random variable X has the *binomial distribution with parameters n and p* if X has a discrete distribution for which the probability function as follows

$$f(x|n,p) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{for } x = 0, 1, 2, 3..., n, \\ 0 & \text{otherwise} \end{cases}$$

1 Lecture Note(10.19)

1.1 Mean and Variance(Scaling Formula)

Mean for the linear function can :
$\mathbb{E}(g(X)) = g(\mathbb{E}(X)) \quad g(X) = aX + b$
$\mathbb{E}(b + \sum_{i=0}^{n} a_i * X_i) = b + \sum_{i=0}^{n} a_i * \mathbb{E}(X_i)$

$$Var(X) = \mathbb{E}X^2 - \mu_x^2 \mu_X^2 \text{ where } \mu_X = \mathbb{E}X$$

Now one can derive the scaling formula for the variance of a linear transformation of a random variable using the same kind of argument as above:

$$Var(aX + b) = a^2 Var(X)$$
$$\sigma_{aX+b} = a \cdot \sigma_x$$

The variance of a linear combination (i.e. a sum) of random variable is more complicated that the expectation of a linear combination of random variables; we'll get to that later.

As a simple example , suppose we have some probability distribution for a temperature measured in Fahrenheit, called X. Let's suppose we have reported its mean and standard deviation. if we want the mean and standard deviation of that temperature measured in Centigrade

$$Y = \frac{5}{9}(X - 32) = \frac{5}{9}X - \frac{160}{9}$$
$$\mathbb{E}Y = \frac{5}{9}\mathbb{E}X - \frac{160}{9}$$
$$\sigma_Y = \frac{5}{9}\sigma_X$$

****Figure 1.1**
Note this corresponds to mapping the "confidence interval"$[\mu_X - \sigma_X, \mu_X + \sigma_X]$ by simply mapping the endpoints under the linear transformation.

$$[\mu_Y - \sigma_Y, \mu_Y + \sigma_Y]$$

1.2 Moment generating function

This is a very useful trick in a variety of probability and stochastic models, particular in more advanced settings. In this class, we will just illustrate it on some basic examples. The reason that the moment generating function can simplify probability calculations is the same way as a Laplace transform or a Fourier transform or a Z-transform can simplify calculations of certain problems. They work when the problem has a certain symmetry or structure that plays well with transform.

This will come out in later lecture, for now let's just take a pedestrian approach and see how the moment generating function works.
Fundamental property that gives the moment generating function(mgf)

$$M_X(s) = \mathbb{E}e^{sX}$$

(This is like the Laplace transformation of the pmf) s is a given number,(eg: 1,3 π), no meaning. sX is a real value. its name is this:

$$\mathbb{E}X^m(m^{th}momentof\bar{X}) = \left(\frac{d}{ds}\right)^m M_x(s)\Big|_{s=0} \text{ for any } m = 0, 1, 2, \cdots$$

$\mathbb{E}X^m$ is the mth moment of X (Easily to calculate derivative than integral.) The first and second moment are important for computing mean and standard deviation.
Why does the moment generating property work?

For $m = 1 : \frac{d}{ds}M_x(s) = \frac{d}{ds}\mathbb{E}e^{sX} = \mathbb{E}\frac{d}{ds}e^{sX} = \mathbb{E}Xe^{sX}$
**meaning: $\frac{d}{dx}$: deterministic linear operation (derivative of a sum = sum of a derivative)
**(but have to do some careful justification if the sum is infinite in order to exchange derivatives and infinite sums; but this does work provided s is sufficiently small.)

$$\left(\frac{d}{ds}\right)^m M_x(s)\Big|_{s=0} M_x(s) = \mathbb{E}Xe^{9X} = \mathbb{E}X$$

Now let's compute mean and standard deviation for come of the basic discrete random variable models:

Discrete uniform distribution

$$R(x) = \{x_1, x_2, .., x_k\}$$

$$p_x = \frac{1}{k} \text{for } x \in R(X)$$

$$\mu_X = \mathbb{E}X = \sum_{x \in R(X)} xp_x = \frac{1}{k}\sum_{i=1}^{k} x_i$$

$$Var(X) = \mathbb{E}((X - \mu_X)^2) = \sum_{(X \in R(X)} (x - \mu_x)^2 p_x = \frac{1}{k}\sum_{i=1}^{k}(x_i - \mu_x)^2$$

$$\sigma_X = \sqrt{\frac{1}{k}\sum_{i=1}^{k}(X_i - \mu_X)^2}$$

(2) Binomial distribution(Bernoulli is a special case)

$$R(x) = \{0, 1, 2, \cdots, n\}, p_x = \binom{n}{x} p^x q^{n-x} \text{ where } q = 1 - p$$

$$\mathbb{E}X = \sum_{x \in R(x)} x p_x = \sum_{x=0} x \binom{n}{x} p^x q^{n-x}$$

This can be evaluated (as can the second moment, but it requires clever manipulations of the combinatorial coefficient)
Two cleaner ways to compute the mean and SD other than this discrete approach

- View the binomial random variable as a sum of n Bernoulli random variables; we will talk more about that later.
- Moment generating function:

$$\begin{aligned} M_X(s) = \mathbb{E}e^{sX} &= \sum_{x \in R(X)} e^{sX} p_x = \sum_{x=0}^{n} e^{sX} \binom{n}{x} p^x q^{n-x} = \sum_{x=0}^{n} \binom{n}{x} (pe^s)^x q^{n-x} \\ &= (pe^s + q)^n \end{aligned}$$

So now:

$$M'_X(s) = n(pe^s + q)^{n-1} pe^s$$

$$\begin{aligned} M''_X(s) &= n(n-1)(pe^s + q)^{n-2}(pe^s)^2 + n(pe^s + q)^{n-1} pe^s \\ &= n(pe^s + q)^{n-2} pe^s((n-1)pe^s + pe^s + q) \\ &= n(pe^s + q)^{n-2} pe^s(npe^s + q) \end{aligned}$$

So the expected value for binomial distribution :

$$\mathbb{E}X = M'_x(0) = n(pe^0 + q)^{n-1} pe^0 = n(p+q)^{n-1} p = np$$

$$\mathbb{E}X^2 = M''_X(0) = n(pe^0 + q)^{n-2} pe^0(npe^0 + q) = (np)^2 + npq - (np)^2$$

$$\begin{aligned} Var(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 &= np(np + q) - (np)^2 \\ &= (np)^2 + npq - (np)^2 \end{aligned}$$

$$Var(X) = npq$$

$$\sigma_x = \sqrt{npq}$$

Image null hypothesis model where every voter was completely undecided and flopped a coin to decide who they'd vote for. The average number of votes for a given candidate can be, under this null hypothesis model, given by a random variable which is binomially distributed with n=$6 * 10^6$ and $p = \frac{1}{2}$ (Bush won Florida by 537 votes) the standard deviation will be

$\sigma_X = \sqrt{npq} = \sqrt{n \times \frac{1}{2} \times \frac{1}{2}} = \frac{1}{2}\sqrt{n} \approx 1200$
So within interval $[\mu_X - \sigma_X, \mu_X + \sigma_X] = 3 * 10^6 \pm 1200$. Which is ...
3) Hyper-geometric distribution? It/s a pain, we'll deal with it later. Mgf doesn't help here.
4) Poisson Distribution

$$R(X) = \{0, 1, 2, ...\}, p_x = \frac{e^{-\lambda}\lambda^x}{x!}$$

It/s not so hard to compute the sums for the mean and second moment directly but for practice, we'll use mgf

$$\begin{aligned} M_X(s) &= \mathbb{E}e^{sX} \\ &= \sum_{x=0}^{\inf} e^{sx}\frac{e^{-\lambda}\lambda^x}{x!} \\ &= e^{-\lambda}\sum_{x=0}^{\inf} e^{sx}\frac{\lambda^x}{x!} \\ &= e^{-\lambda}e^{e^s\lambda} = e^{-\lambda+e^s\lambda} \\ &= e^{\lambda(e^s-1)} \end{aligned}$$

To compute moments:

$$M_X'(s) = \lambda e^s e^{\lambda(e^s-1)}$$

$$M_X''(s) = \lambda e^s e^{\lambda(e^s-1)} + \lambda e^s \lambda e^s e^{\lambda(e^s-1)} = e^{\lambda(e^s-1)}\lambda e^s(1 + \lambda e^s)$$

$$\mathbb{E}X = M_X'(0) = \lambda e^0 e^{\lambda(e^0-1)} = \lambda * 1 * 1$$
$$\mathbb{E}X = \lambda$$

$$\mathbb{E}X^2 = M_X''(0) = e^{\lambda(e^0-1)}\lambda e^0(1 + \lambda e^0) = 1 \times \lambda \times 1(1 + \lambda) = \lambda(1 + \lambda)$$

$$Var(X) = \mathbb{E}X^2 - (\mathbb{E}X)^2 = \lambda(1 + \lambda) - \lambda^2 = \lambda + \lambda^2 - \lambda^2 = \lambda$$
$$\sigma_X = \sqrt{Var(X)}$$

So this affirms what we alluded when we defined the Poisson distribution, that the parameter λ is just the mean of the random variable. Note also that the standard deviation scales as the square root of the mean, similarly to binomial distribution
Negative binomial distribution(geometric distribution is a special case)

$$R(X) = 0, 1, 2...p_x = \binom{x + r - 1}{x} p^r q^x = \binom{-r}{x} p^r (-q)^x$$

Mean and standard deviation doable with pain by direct sum. Let's use mgfs.

$$
\begin{aligned}
M_X(s) = \mathbb{E}e^{sX} &= \sum_{x=0}^{\infty} e^{sx}\binom{-r}{x}p^r(-q)^x \\
&= \sum_{x=0}^{\infty} e^{sx}\binom{-r}{x}p^r(-qe^s)^x \\
&= p^r\sum_{x=0}^{\infty}\binom{-r}{x}(-qe^s)^x 1^{-r-x} = p^r(-qe^s+1)^{-r} \\
M_X(s) &= p^r(-qe^s+1)^{-r} \\
M_X'(s) &= p^r(-r)(-qe^s+1)^{-r-1}(-qe^s) \\
M_x''(s) &= p^r(-r)(-r-1)(-qe^s)^2(-qe^s+1)^{-r-2} + p^r(-r)(-qe^s+1)^{-r-1}(-qe^s) \\
&= p^r(-r)(-qe^s)(-qe^s+1)^{-r-2}[(-r-1)(-qe^s)+(qe^s+1)] \\
&= p^r(-r)(-qe^s)(-qe^s+1)^{-r-2}[qre^s+1]
\end{aligned}
$$

$$
\mathbb{X} = M_X'(0) = p^r(-r)(-qe^0+1)^{-r-1}(-qe^0) = p^r(-r)(1-q)^{-r-1}(-q) = p^r(-r)(p)^{-r-1}(-q)
$$

$$
\begin{aligned}
\mathbb{E}X &= \frac{qr}{p} \\
\mathbb{E}X^2 &= M_X''(0) \\
&= p^r(-r)(-qe^0)(-qe^0+1)^{-r-2}[qre^0+1] \\
&= p^r rq(1-q)^{-r-2}(qr+1) \\
&= p^r qrp^{-r-2}(qr+1) = \\
&= \frac{qr(qr+1)}{p^2} \\
Var(X) &= \mathbb{E}x^2 - (\mathbb{E}x)^2 \\
&= \frac{qr(qr+1)}{p^2} - (\frac{qr}{p})^2 \\
&= \frac{(qr)^2}{p^2} + \frac{qr}{p^2} - \frac{(qr)^2}{p^2} \\
Var(X) &= \frac{qr}{p^2} \\
\sigma_x &= \frac{\sqrt{qr}}{p}
\end{aligned}
$$

2 Lecture Note(10.23)

Application of discrete random variables and mgfs to a family planning model.

Let's consider three strategies by which families may be formed

1. Strategy A: every family has exactly c children per couple
2. Strategy B: Every family keeps having children until they have 1 child with special feature(CSF), then they stop.
3. Strategy c every family keeps having children until they have 2 CFSs, they they stop.

Assumpe each child is a CSF with probability p, independently of other children. Consider how the distribution of CSFs in population depend on these family planning strategies.
1. What is the ratio of CSF to children in the population as a whole?(sampling children).
2. What is the average ratio of CSF to children within a typical family?(sampling families)

the mathematical setup for these questions are, respectively:

1. Ratio of CSFs to children in the population as a whole: let's define X_i to be the number of CSFs in the i^{th} family in the population. Then the number of CSFs in population is $\sum_{i=1}^{n}$ if we have n families.
   The total number of children is $\sum_{i=1}^{n} C_i$ is the number of children in family i
   So the ratio of CSFs to children in the population:

$$\frac{\sum_{i=1}^{n} X_i}{\sum_{i=1}^{n} C_i} = \frac{\frac{1}{n}\sum_{i=1}^{n} X_i}{\frac{1}{n}\sum_{i=1}^{n} C_i} =_{n\to\infty} \frac{\mathbb{E}X}{\mathbb{E}C}$$

   By the frequentist interpretation of expectation, or by law of large Numbers

2. The answer to the second question is instead:

$$\frac{1}{n}\sum_{i=1}^{n} \frac{X_i}{C_i} \to_{n\to\infty} \mathbb{E}\frac{X}{C}$$

   (This is not the same as the equation in the 1)
   Strategy A: C= c is deterministic constant.

$$X \; bin(c,p) \text{(binomially distributed with c trials and success probability p)}$$

$$\mathbb{E}C = c$$
$$\mathbb{E}X = cp$$
$$\mathbb{E}\left(\frac{X}{C}\right) = \frac{1}{c}\mathbb{E}X = \frac{cp}{c} = p$$
$$\frac{\mathbb{E}X}{\mathbb{E}C} = \frac{cp}{c} = p$$

Strategy B: $X = 1$

$$C = 1 + N \text{where N is the number of non-special children}$$

The random variable N is a geometric distribution with success probability p The pmf for N is $p_n = pq^n$ for $n = 0, 1, 2, \cdots$
where $q = 1 - p$

$$\mathbb{E}X = \mathbb{E}1 = 1$$
$$\mathbb{E}C = \mathbb{E}(1 + N) = 1 + \mathbb{E}N = 1 + \frac{q}{p} = \frac{p+q}{p} = \frac{1}{p}$$

That we could also have gotten directly by using the geometric distribution for total number of trails unt
Therefore, the fraction of special children in the population is

$$\frac{\mathbb{E}X}{\mathbb{E}C} = \frac{1}{\frac{1}{p}} = p$$

Now what is the average fraction of special children within a family?

$$\begin{aligned}\mathbb{E}\left(\frac{X}{C}\right) &= \mathbb{E}\left(\frac{1}{1+N}\right) \\ &= \sum_{n=0}^{\infty} \frac{1}{1+n} p_n \\ &= \sum_{n=0}^{\infty} \frac{1}{1+n} pq^n\end{aligned}$$

Try to make this look like a moment calculation so we can use mgf approach.

$$c = n + 1 \text{(index of total number of children):}$$

$$\mathbb{E}(\frac{X}{C}) = \sum_{c=1}^{\infty} \frac{1}{c} pq^{c-1}, pc^{c-1} \text{ is pmf of C}$$

This is exactly $\mathbb{E}C^{-1}$ Let's compute the mgf of C:

$$\begin{aligned}\mathbb{E}e^{sC} &= \mathbb{E}^{s(N+1)} \\ &= \mathbb{E}e^{sN+s} \\ &= e^{sN}e^{s} \\ &= e^{s}\mathbb{E}e^{sN} \\ &= e^{s}M_N(s)\end{aligned}$$

But N is geometric with success probability p, so we know its mgf from last time: $M_N(s) = p(1 - qe^s)^{-1}$

$$M_C(s) = \frac{pe^s}{1 - qe^s}$$

What we want is $\mathbb{E}C^{-1}$. Let's try integrating the MGF rather than differentialing it:

$$\begin{aligned}\int_{-\infty}^{0} M_C(s)ds &= \mathbb{E}\int_{-\infty}^{0} e^{sC}ds \\ &= \mathbb{E}\frac{e^{sC}}{C}|_{s=-\infty}^{0} \\ &= \mathbb{E}\left(\frac{e^{0C}}{C} - 0\right) \\ &= \mathbb{E}\frac{1}{C}\end{aligned}$$

So therefore

$$\begin{aligned}\mathbb{E}\frac{1}{C} &= \int_{-\infty}^{0} \frac{pe^S}{1-qe^s}ds \\ &= \int_0^1 \frac{pdu}{1-qu} \\ &= -\frac{p}{q}\ln(|1-qu|)|_0^1 \\ &= -\frac{p}{q}(\ln(|1-q|) - \ln(1)) \\ &= -\frac{p}{q}(\ln(1-q))\end{aligned}$$

$$\text{Therefore}\mathbb{E}(\frac{X}{C}) = \frac{p}{q}\ln(\frac{1}{1-q})$$

is the average fraction of a family's children that are special.

Continuous variable

Examples of continuously distributed random variable:

- uncertain spatial locations (...)
- uncertain times(length of an illness, time at which a person infects another person, lifetime of a piece of equipment or length of time between failures of equipment)
- measuring various physical variable (momenta, temperature,velocities)
- some situations where the natural units are discrete, a continuous model is
  - dynamics of large populations
  - financial markets.

Why use continuous models as approximations to discrete models? if you keep a discrete model, then the model often has an extra parameter which is the discretization unit, and that can actually complicate the analysis.
A key distinction between continuous and discrete random variable is that probability mass function are useless for continuous random variables because generally speaking (not always, but typically)

$$Pr(X = x) = 0$$

Therefore the probability mass function does not do any useful work for continuous random variables
To extend the framework of random variable theory to account for general random variables (that are not discrete) appeals to measure theory. We will not go into this, but rather focus just on the special case of what's known as (absolutely) continuous random variables. For now we'll focus just on continuous random variables which have a one-dimensional state space which is real:$S \subseteq \mathbb{R}$. Such random variables have the property that there exists a probability density function(pdf)(which replaces the pmf) (use f(x) to represent pdf, f_x represents pmf)

$$f(x)(\in L^1) \text{ one dimension}$$

Which has the properties:

- $f(x) \geq 0$
- $\int_S f(x)dx = 1$

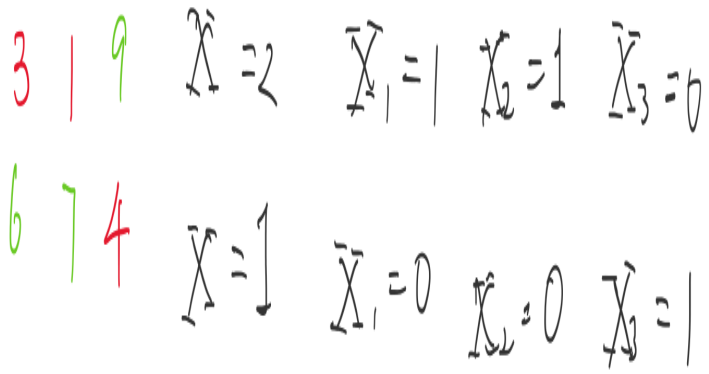
Such that for any nice (Borel) subset $B \subseteq S$, we have

$$Pr(X \in B) = \int_B f(x)dx$$

In particular, if we consider B to be an interval between values a,b and it doesn't matter if we consider it an open or closed or half-open interval:

$$Pr(a < X < b) = Pr(a \leq X \leq b) = Pr(a < X \leq b) == Pr(a \leq X < b)$$

$$= \int_a^b f(x)dx$$



(figure 2.1)
Notice that when we talk about continuous random variables, we need to talk about probabilities for the random variables to fall within some nice set, rather than some particular point, because the probability of the latter is always zero. What's the intuitive meaning of the pdf? Note that while it plays an analogous role to the probability mass function of discrete random variables, it's not the same thing, and it doesn't even have units of probability

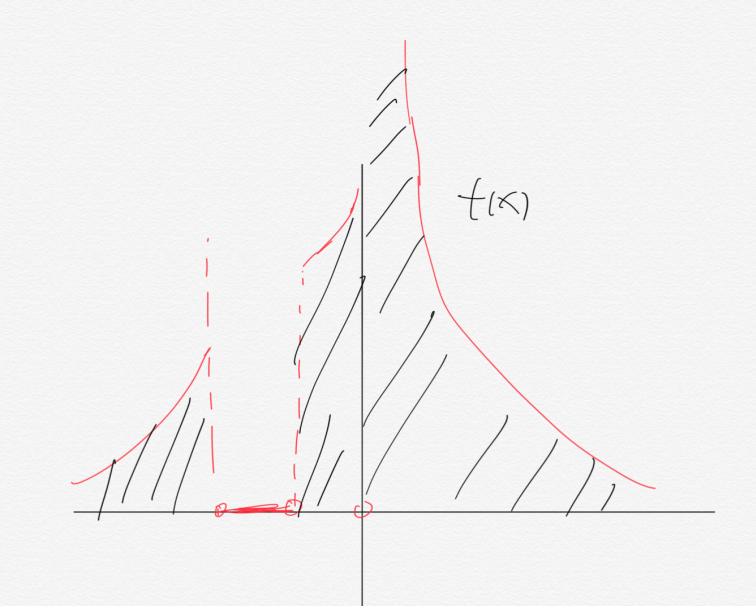
It is not true $f(x) = Pr(X = x)$ No it is not

$$Pr(a < X < b)\text{dimensionless} = \int_a^b f(x)dx[\text{f}][\text{x}]$$

So the dimensions of the PDF $[f] = \frac{1}{[x]}$
So if X has units of length, then f ahs units of 1/length
If X has units of time, the f has unites of $\frac{1}{time}$
The probability density function plays an exactly analogous role to the mass density function in associating a region to a probability/mass by integrating the density function over that region.
So what does the pdf mean? consider the following set of equatlities which are valid where f(x) is continuous(which is not necessary!):

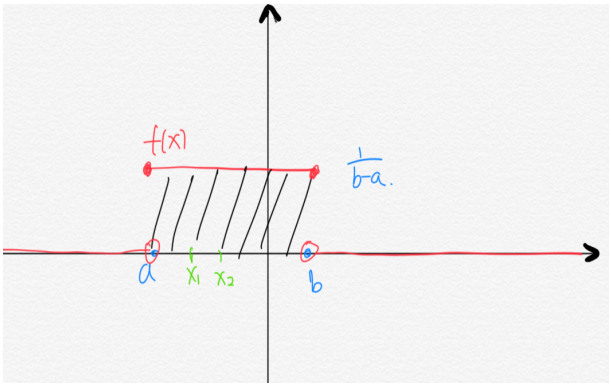
$$\begin{aligned} f(x) &= \lim_{\epsilon \to 0} \frac{\int_{x-\epsilon}^{x+\epsilon} f(x')dx'}{2\epsilon} (\text{Mean Value Theorem}) \\ &= \lim_{\epsilon \to 0} \frac{\Pr(x - \epsilon < X < x + \epsilon)}{2\epsilon} \\ &= \lim_{\epsilon \to 0} \frac{\Pr(|X - x| < \epsilon)}{2\epsilon} \end{aligned}$$

That said note that probability density function don't have to be bounded or continuous.



Now let's look at some important continuous random variable models.
Uniform(continuous) distribution on an interval $[a, b]$(a¡b)

$$X \sim U(a, b)$$



We've alluded to this probability distribution before (when we did Bertrand experiment, etc.)

Want, for any $(x_1, x_2) \subseteq [a, b]$, we want $Pr(x_1 \leq X \leq x_2) = \frac{|x_2 - x_1|}{b-a} = \int_{x_1}^{x_2} f(x)dx$
What f(x) works? Well the idea of uniform distribution suggests f(x) should be some constant between a and b. But the area under f(x) should be 1,
We could just define the pdf to be $f(x) = \frac{1}{b-a}$ on the state space $S = [a, b]$
Sometimes though, people like to take the state space of any real random variable as $S = \mathbb{R}$
And then define pdf

$$f(x) = \begin{cases} 0 & x < a \\ \frac{1}{b-a} & a \leq x \leq b \\ 0 & x > b \end{cases}$$

Sometimes continuous random variables are uniformly distributed.

3 Lecture Note(10.26)

Last time we introduced the concept of probability density function(pdf) that plays the role for continuous random variable that probability mass function (pmf) did for discrete random variables.

3.1 From Discrete Random Variables to Continuous Random Variables

How do we carry over the calculations for discrete random variables to continuous random variables?

- Expectations?
- Expectations of $g(X)$
- Compute probability distribution of $Y = g(X)$ from the probability distribution of X(next time)

One defines the expectation of a continuous random variable X with pdf $f(x)$ and state space(range) S as:

$$\mu_X = \mathbb{E}X = \langle X \rangle = \int_S x f(x) dx$$

(note the analogy to expectation of discrete random variables; use pdf rather than pmf and integrate rather than sum.) This definition can be shown through measure theory to be unified naturally with the definition of expectation for discrete random variables(as we'll see in the next lecture theory), or one can see that if the continuous random variable will have approximately the expectation of the approximating discrete random variable (Riemann sum argument.)

Then the LUS for continuous random variables:

$$\mathbb{E}_g(X) = \int_S g(x) f(x) dx$$

Linearity of expectation formula applies to all random variables, not just discrete random variables. So do the scaling formulas for expectations and standard deviation.
Let's practice these formulas with the continuous uniform distribution we introduced last time.

Let's start by studying the mean and standard deviation of $U \sim U(0,1)$ a uniformly distribution random number on [0,1]

$$f_U(x) = \text{(subscript to describe the random variable)} \begin{cases} 1 & \text{for } 0 \leq x \leq 1 \\ 0 & \text{else} \end{cases}$$

$$\mathbb{E}U = \int_S x f_U(x) g x = \int_0^1 x 1 ds(+0) = \int_0^1 x dx = \frac{1}{2} x^2 \Big|_0^1 = \frac{1}{2}(1-0) = \frac{1}{2}$$

second moment?

$$\mathbb{E}U^2 = \int_S x^2 f_U(x) g x = \int_0^1 x^2 1 ds(+0) = \int_0^1 x^3 dx = \frac{1}{3}(1-0) = \frac{1}{3}$$

$$VarU = \mathbb{E}U^2 - (\mathbb{E}U)^2 = \frac{1}{3} - (\frac{1}{2})^2 = \frac{1}{12}$$

$$\sigma_U = \sqrt{Var\ U} = \frac{1}{\sqrt{12}}$$

What about calculating these statistics for $X \sim a + (b-a)U$.
Could redo the calculation. Or let's exploit the relationship:

$$X \sim a + (b-a)U.$$

This statement with the $\sim$ means both sides have the same probability distribution, but might not be exactly equal. If X_1 and X_2 are the Bernoulli random variable describing the outcome of trials 1 and 2 in a sequence od Bernoulli trials, then $X_1 \neq X_2$
But $X_1 \sim X_2$ (they have same probability distribution but not same random variable)
How do we prove that $X \sim a+(b-a)U$? We will develop a machine for doing this next time, but let's do a basic version today to demonstrate the idea.

We want to show that for any $x_1 < x_2$ that $P(x_1 < X < x_2) = P(x_1 < a + (b-a)U < x_2)$
We can check that if $x_1 < x_2 < a$ or $b < x_1 < x_2$ then both probabilities are 0. How?
Manipulate the right hand side by isolating U.

$$P(x_1 < a+(b-a)U < x_2) \xrightarrow{-a} P(x_1-a < (b-a)U < x_2-a) \xrightarrow{/(b-a)} P(\frac{x_1-a}{b-a} < U < \frac{x_2-a}{b-a})$$

What about $a < x_1 < x_2 < b$? Then

$$\begin{aligned} P(x_1 < X < x_2) &= \int_{x_1}^{x_2} f_x(x)dx \\ &= \int_{x_1}^{x_2} \frac{1}{b-a}dx \\ &= \frac{x_2-x_1}{b-a} \\ P(x_1 < a + (b-a)U < x_2) &= P(\frac{x_1-a}{b-a} < U < \frac{x_2-a}{b-a}) \\ &= \frac{x_2-a}{b-a} - \frac{x_1-a}{b-a} \\ &= \frac{x_2-x_1}{b-a} \end{aligned}$$

Because $0 \leq \frac{x_2-a}{b-a} \leq \frac{x_1-a}{b-a} \leq 1$
We see both probabilities are the same. And then one can check the cases where the interval$[x_1, x_2]$ only partially overlaps the interval $[a, b]$, again the Probabilities agree. We have sketched how to show that $X, a + (b-a)U$ have the same probabilities to fall in any interval we specify. Intuitively, and next time rigorously, this is enough to show that they have the same probability distribution:$X \sim a + (b-a)U$
So what? Note that this means that X has the same probability distribution as linear function $a + (b-a)U$ of the random variable U

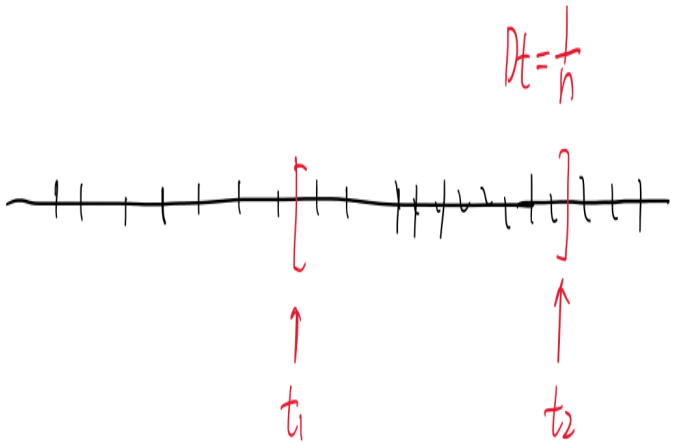
SO

$$\begin{aligned}
\mathbb{E}X &= (a + (b-a)U) \\
&= a + \frac{(b-a)}{2} \\
&= a + (b-a)\frac{1}{2} \\
&= a + b\frac{1}{2} - a\frac{1}{2} \\
&= \frac{a+b}{2} \\
\sigma_x &= \sigma_{a+(b-a)U} \\
&= (b-a)\sigma_U \\
&= (b-a)\frac{1}{\sqrt{12}} \\
&= \frac{b-a}{\sqrt{12}}
\end{aligned}$$

3.2 Poisson Point Process

We'll begin building the bridge between Bernoulli trials and continuous time randomness by considering a continuum limit of Bernoulli trials

Bernoulli process in the $n \to \infty$



The number of success in $[t_1, t_2]$ is a random variable, let's call it N, it's a discrete random variable. The only tricky thing is how to handle success that fall in the small intervals that contain the endpoint t_1, t_2 but since the probability for that happen will go to zero as $n \to \infty$ we will neglect that. Then the number of successes that fall within one of the Bernoulli trials contained in the interval $[t_1, t_2]$ can be viewed as the number of successes in $\frac{t_2 - t_1}{\frac{1}{n}} \pm 2$ trials with success probability $p_n = \frac{1}{n\tau}$ We'd get a binomial distribution with $n(t_2 - t_1) \pm 2$ trials

and success probability $p_n = \frac{1}{n\tau}$ As $n \to \infty$ this converges to a Poisson distribution with mean

$$\begin{aligned}\lambda &= \lim_{n\to\infty}(n(t_2 - t_1) \pm 2)\frac{1}{n\tau} \\ &= \lim_{n\to\infty}(\frac{(t_2 - t_1)}{\tau} \pm \frac{2}{n\tau}) \\ &= \frac{t_2 - t_1}{\tau}\end{aligned}$$

In a Poisson point process with parameter τ (which is the continuum limit of a Bernoulli process), The number if success in an interval $[t_1, t_2]$ is Poisson distributed with mean $\lambda = \frac{t_2 - t_1}{\tau}$

Now look at the **Gamma experiment** for another view. How long do we have to wait until the first success, as $n \to \infty$
We know the answer for the number of trials until the first success is given by a geometric distribution, and we want to take the $n \to \infty$, but this is going to give a continuous probability distribution. How do we take limits of discrete random variables to get continuous random variables?
Need a framework that unifies both continuous and discrete random variables, and as we'll see next time, the cumulative distribution function (CDF) is the basic tool for unifying essentially all practical random variables.
Today we will just show the idea; make it systematic next time.
Let X_1 denote the index of the first trial at which a success occurs. And let $T_1 = \frac{X_1}{n}$ define the time at which the success occurs(since each trial takes a time $\frac{1}{n}$).
$\Pr(T_1 > t)$ the probability that the first success takes at least until some time $t > 0$

$$\begin{aligned}Pr(T_1 > t) &= Pr(\frac{X_1}{n} > t) = Pr(X_1 > nt) = Pr(X_1 > \lfloor nt \rfloor) \\ &= (1 - p_n)^{\lfloor nt \rfloor}\end{aligned}$$

$\lfloor y \rfloor$ is the greatest integer not exceeding y
Because requiring more than $\lfloor nt \rfloor$ trials to get the first success means the first $\lfloor nt \rfloor$ trials were all failures.

$$\begin{aligned}\lim_{n\to\infty} &= \left(1 - \frac{1}{n\tau}\right)^{\lfloor nt \rfloor} \\ &= \lim_{n\to\infty}\left[\left(1 - \frac{1}{n\tau}\right)^{-n\tau}\right]^{\frac{-\lfloor nt \rfloor}{n\tau}} \\ &= e^{-\frac{t}{\tau}}\end{aligned}$$

(compound interest rate formula)
That

$$\lim_{h\to 0}(1 + h)^{\frac{1}{h}} = e$$

so that implies, by taking $h = -\frac{1}{n\tau}$ that $\lim_{n\to\infty}(1 - \frac{1}{n\tau})^{-n\tau} = e$

$$\begin{aligned}\lim_{n\to\infty}\frac{-\lfloor nt\rfloor}{n\tau} &= \lim_{n\to\infty}\frac{-nt}{n\tau}+\frac{nt-\lfloor nt\rfloor}{n\tau}\\ &= \lim_{n\to\infty}-\frac{t}{\tau}+O(\frac{1}{n})\\ &= -\frac{t}{\tau}\end{aligned}$$

$$\Pr(T_1 > t) = e^{-\frac{t}{\tau}}$$

But by the definiton of pdf:

$$\Pr(T_1 > t) \ = \int_t^\infty f_{T_1}(t')dt' = e^{-\frac{t}{\tau}} \text{ for } t > 0$$

Differentiate both sides:

$$\frac{d}{dt}\int_t^\infty f_{T_1}(t')dt' = \frac{d}{dt}e^{-\frac{t}{\tau}}$$

FTC:

$$-f_{T_1}(t) = -\frac{1}{\tau}e^{-\frac{t}{\tau}} \text{ for } t > 0$$

Obviously the PDF is zero by definition of T_1 for $t < 0$

So now we have that the time to wait until the first success in a poisson point process is given by an **exponential distribution** with parameter τ
PDF:

$$f_{T1}(t) = \begin{cases}\frac{1}{\tau}e^{-\frac{t}{\tau}} \text{ for } t \geq 0\\ 0 \text{ for } t < 0\end{cases}$$

This is the continuous analog(for Poisson point processes) to the geometric distribution that describes the number of failures until the first success. The continuous analog to negative binomial distribution to wait for r success in Bernoulli trials is the Gamma distribution for Poisson point process.
Let's compute the basic statistics of the exponential distribution. For practice, let's use mgfs.

$$\begin{aligned}M_{T_1}(s) &= \mathbb{E}E^{sT_1}\\ &= \int_0^\infty e^{st}f_{T1}(t)dt\\ &= \int_0^\infty e^{sT}\frac{1}{\tau}e^{-\frac{t}{\tau}}dt\\ &= \frac{1}{\tau}\int_0^\infty e^{t(s-\frac{1}{\tau})}dt\\ &= \frac{1}{\tau}\frac{1}{s-\frac{1}{\tau}}e^{t(s-\frac{1}{\tau})}\Big|_{t=0}^{\infty}\\ &= \frac{1}{\tau}\frac{1}{s-\frac{1}{\tau}}(0-1) \text{ providrd that } s < \frac{1}{\tau}\\ M_{T_1}(s) &= -\frac{1}{\tau\left(s-\frac{1}{\tau}\right)}\\ &= \frac{1}{1-s\tau}\end{aligned}$$

So now:

$$\begin{aligned}
\mathbb{E}T_1 &= \frac{d}{ds} M_{T_1}(s)\Big|_{s=0} \\
&= -(-\tau)\frac{1}{(1-s\tau)^2}|_{s=0} \\
&= \tau \\
\mathbb{E}T_1^2 &= \frac{d^2}{ds^2} M_{T_1}(s)\Big|_{s=0} \\
&= 2\tau^2 \\
Var(T_1) &= \mathbb{E}T_1^2 - (\mathbb{E}T_1)^2 \\
&= 2\tau^2 - \tau^2 = \tau \\
\sigma_{T_1} &= \tau
\end{aligned}$$

This is the continuous random variable analog to the geometric distribution for discrete random variables. One can show that the gamma distribution
It doesn't make sense any more, after the continuum limit, to list failures and successes. Instead simply label the sequence of times at which an incident or a success happens. We could have done this for the Bernoulli process.

The times Y_j at which the incidents occur are separated by interincident times T_j so $Y_n = \sum_{j=1} T_j$ Yhe random variables T_j are easier to work with.

Poisson process is useful for modeling various continuous time events where something special occurs at discrete moments of time:

- arrival of demand or request
- entry of vehicles into a roadway
- arrival of pedestrians at a crosswalk
- moments at which a one-step biochemical reaction occurs
- simple models for when a neuron receives a signal from a neighboring neuron
- price shocks in markets

This is the continuous random variable

4 Lecture Note(10.30)

4.1 Normal (Gaussian) Distribution

One writes as shorthand: $X \sim N(\mu, \sigma^2)$ to refer to a normal (Gaussian) PDF of the form:

$$f(x) = \frac{e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}}{\sqrt{2\pi}\sigma}$$

(figure 10.30.1)
Let's compute its mean and variance using a similar strategy to what we did for the uniform distribution.
Standard Gaussian normal random variable: $Z \sim N(0, 1)$ (centered at 0, width 1) Once can show, by the method we will develop below, that $x \sim N(\mu, \sigma^2)$ can be expressed $X \sim \sigma Z + \mu$

By doing some standard Gaussian integrals, one finds that for the standard Gaussian random variable:

$$\mathbb{E}Z = 0$$

$$\sigma_z = 1$$

We can then use there results to conclude:
If $X \sim N(\mu, \sigma^2)$ then

$$\mathbb{E}X = \mathbb{E}(\sigma Z + \mu) = \sigma \mathbb{E}Z + \mu = \sigma \cdot 0 + \mu = \mu$$

$$\sigma_x = \sigma_{\sigma Z + \mu}$$

Gaussian distributions are good models for many variables, particularly those which result from a large number of independent factors, by virtue of the Central Limit Theorem, which we'll talk about later. For now, observe that Gaussian/normal distribution is a good approximation to :

- binomial distribution when the number of trial is large, and the success probability is not necessarily small (match mean and variance), the mean must be large for the approximation to be good (n and np greater enough)
- Poisson distribution when its mean is large(match mean and variance) ($\lambda$is greater enough)

This doesn't simplify the number of parameters, but complex computations with Gaussian are in general considerably easier than with binomials or Poisson. The general idea is that multiple Gaussian random variables can be analyzed via linear algebra.

4.2 Cumulative Distribution Function

Some books call it simply the "distribution function" but this is not common language in my experience and potentially confusing.
The CDF for a random variable X (does not matter if it's discrete, continuous, or neither)

$$F_X(x) = Pr(X \leq x)$$

It is a unified way to represent random variables with complete information. Let's see how it's related to the previous ways of describing discrete and continuous random variables

For discrete random variables X

$$F_X(x) = \sum_{y \leq x, y \in R(x)} p_y \text{ where } p_x \text{ is the pmf of } X.$$

The pmf for a discrete random variable can be obtained from the CDF by looking at the location and size of jumps.
For continuous random variable X image here

$$\begin{aligned} F_X(x) &= \Pr(X \in (-\infty, x]) \\ &= \int_{-\infty}^{x} f_X(y)dy \text{ where } f_X(x) \text{ is the pdf of } X \end{aligned}$$

Conversely, we can get the PDF from the CDF by differentiation:

$$f_x(x) = \frac{d}{dx} F_x(x)$$

Let's illustrate for exponential distribution

$$\begin{aligned} \text{PDF: } f(x) &= \begin{cases} \frac{1}{\tau} e^{-\frac{x}{\tau}} \text{ for } x \leq 0 \\ 0 \text{ for } x < 0 \end{cases} \\ \text{CDF:} F(x) &= \int_{-\infty}^{x} f(y)dy = 0 \text{ if } x \geq 0 \\ &= \int_{-\infty}^{0} 0dy + \int_{0}^{x} \frac{1}{\tau} e^{\frac{-y}{\tau}} dy \\ &= 0 - e^{\frac{-y}{\tau}} |_{y=0}^{x} \\ &= -(e^{\frac{-x}{\tau}} - 1) \\ &= 1 - e^{\frac{-x}{\tau}} \end{aligned}$$

Note that discontinuities in PDFs can be handles fairly arbitrarily (whether you set the value =0 to left or right hand limit does not mater)but the discontinuities in CDFs have meaning, and as we will see later ....
We see that CDF

4.2.1 Advantages of CDF

The CDF is an advantageous description in the following sense:

- It's a general purpose description, works for any random variable whether it's discrete, continuous, or neither(hybrid)
- it gives a general purpose way of going from the description of one random variable to the description of a function of that random variable ("derived distribution"). If you have $Y = g(X)$ then it is relatively easy to go from the CDF of X to the CDF of Y But going from the PDF of X to the PDF of Y is more complicates (but doable).
- It plays a central role in a general purpose algorithm for simulating random variables.
- In statistics, CDFs are used as a tool for comparing how close two probability distributions are, i.e., the distance between the distribution of some data and a theoretical model.(Kolmogorow-Smirnov test)

4.2.2 Disadvantages of CDF

But the CDF has the following disadvantages:

- more awkward to work with than the PMF or the PDF when these are well-defined eg: for computing statistics.
- through its definition can be extended to multiple dimensions(for several simultaneous random variables), it's very awkward in more than one dimension.

4.3 Properties and Uses of the CDF.

We now proceed to develop the properties and uses of the CDF.
For any kind of random variable(discrete, continuous, hybride, crazy) a CDF

- $lim_{x\to-\infty}F_X(x) = 0$
- $lim_{x\to\infty}F_X(x) = 1$
- $F_X(x)$ is a non decreasing function of x
- $F_x(x)$ is right-continuous funciton $F(x) = lim_{y\to x} + F_X(y)$

4.3.1 Use CDF to derive probability distributions

CDF method for deriving probability distributions for functions of random variables (derived distributions)
Suppose we have a random variable X with known probability distribution (either PMF, PDF or CDF) and we want to compute the distribution $Y = g(X)$

Remark: If all we want is $\mathbb{E}Y = \mathbb{E}g(y)$, think first about using LUS.

But continuing with the premise that we really want/need the probability distribution for Y to do some calculation, like $P(Y \in B)$
Recall that after discussion of LUS for discrete random variables, we described a procedure for doing this "change of variables " for discrete random variables. That procedure doesn't make sense for continuous random variables, so we will formulate a procedure that works for all random variables (discrete, continuous, neither)
(CDF conversion)

Let's first use this idea to show that if $Z \sim N(0,1)$ and $X \sim N(\mu, \sigma^2)$
Then $X \sim \sigma Z + \mu$. To prove this it is sufficient to show that the CDF of $g(Z) = \sigma Z + \mu$ is the same as the CDF for X.

Call $V \equiv g(Z) = \sigma Z + \mu$

$$F_v(v) = \Pr(V \leq v) = \Pr(g(Z) \leq v = \Pr(\sigma Z + \mu \leq v)$$

To proceed, "solve the inequality" inside the probability for the random variable whose probability distribution you know (Z):

$$\begin{aligned} F_V(v) &= Pr(\sigma Z + \mu \leq v) = Pr(\sigma Z \leq v - \mu) \\ &= \Pr(Z \leq \frac{v-\mu}{\sigma}) \\ &= F_Z(\frac{v-\mu}{\sigma}) \end{aligned}$$

F_Z is "known" Note though that the CDF for a Gaussian random variable is a special n :function

$$\begin{aligned} F_Z(z) &= \int_{-\infty}^{z} f_z(z')dz' \\ &= \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-\frac{z'^2}{2}} dz' \equiv \Phi(z) \text{ this can show in the answer} \end{aligned}$$

which has to be evaluated numerically; it is not expressible in terms of the standard functions from pre-calculus. $\Phi(z)$ is the CDF for the standard normal random variable; it is closely related to the error function erf, and you can use erf instead of Φ if you want iamge4

$$F_V(v) = \phi(\frac{v-\mu}{\sigma})(Z - score)$$

But can we show that this is $F_x(x)$?

$$F_X(x) = \int_{-\infty}^{x} f_x(x')dx' = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x'-\mu)^2}{2\sigma^2})dx'$$

Let's change variables to make this integral looks like in the integral of a standard normal. $z' = \frac{x'-\mu}{\sigma}$ and $dx' = \frac{1}{\sigma}dx'$. So $x' = \mu + \sigma z'$ and $dx' = \sigma dz'$

$$F_x(x) = \int_{-\infty}^{\frac{x-\mu}{\sigma}} \frac{1}{\sqrt{2\pi}} exp(-\frac{z'^2}{2})dx' = \phi(\frac{x-\mu}{\sigma})$$

So we showed:$F_X(x) = \phi(\frac{x-\mu}{\sigma})$

$$F_V v = \Phi(\frac{v-\mu}{\sigma})$$

Same CDF (different dummy argument) so $V \sim X$
So

$$\sigma Z + \mu \sim X \text{ as claimed}$$

Let's illustrate this procedure by a simple modeling question. Suppose some piece of equipment is used in many places in some industrial operation; its working lifetime is a random variable T which we 'll model as exponentially distributed with mean τ.
The industry employs a block-replacement policy for this piece of equipment, meaning that the equipment is replaced when it breaks or after it has been operating for a time τ_r, whichever comes first

Then what are the statistics(probability distribution, mean variance) for the amount of time this piece of equipment is actually in operation? Call this random variable X
$X = \min(T, \tau_r)$ So can think $X = g(T) = \min(T, \tau_r)$
If all you want is $\mathbb{E}X$, or σ_x then you can use LUS:

$$\mathbb{E}X = \mathbb{E}\min(T, \tau_r) = \int_{-\infty}^{\infty} \min(t, \tau_r)\frac{1}{\tau} exp^{-\frac{t}{\tau}} dt = ...$$

where

$$p_T(t) = \begin{cases} \frac{1}{\tau}e^{-\frac{t}{\tau}} & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases}$$

Similarly to get $\mathbb{E}X^2$ set up a similar calculation using LUS, and get σ_x from item the usual way.

But suppose instead we want to ask questions like $P(X \geq 3)$ then may need to work with CDF approach.
Let's try to get the CDF for X:

$$F_X(x) = \Pr(X \leq x) = \Pr(\min(T, \tau_r) \leq x)$$

Solve the expression inside the probability for T. Can do it algebraically or sometimes graphics plot help.
Sometimes when the function g is peicewise defined, it helps to break the calculation into sub cases.
First suppose $x > \tau_r$. Associate a set $B_x = \{t : g(t) \leq x\}$ Then $B_x = \mathbb{R}$
For $x > \tau_r$,

$$F_x(x) = Pr(X \leq X) = Pr(min(T, \tau_r) \leq x) = Pr(T \in B_x) = Pr(T \in \mathbb{R}) = 1$$

but if $x \leq \tau_r$ Associate a set $B_x = t : g(t) \leq x$. Then $B_x = (-\infty, x)$.

$$\begin{aligned} F_X(x) = Pr(X \leq x) &= Pr(\min(T, \tau_r) \leq x) \\ &= Pr(T \in B_x) \\ &= Pr(T \in (-\infty, x)) \\ &= F_T(x) = 1 - e{-x/\tau_r} for x > 0 \\ &= 0 for x \leq 0 \end{aligned}$$

That tells me that CDF for the working lifetime X:

$$F_X(x) = \begin{cases} 0 & \text{for } x \leq 0 \\ 1 - e^{\frac{-x}{\tau_r}} & \text{for } 0 < x < \tau_r \\ 1 \text{ for } x \geq \tau_r & \end{cases}$$

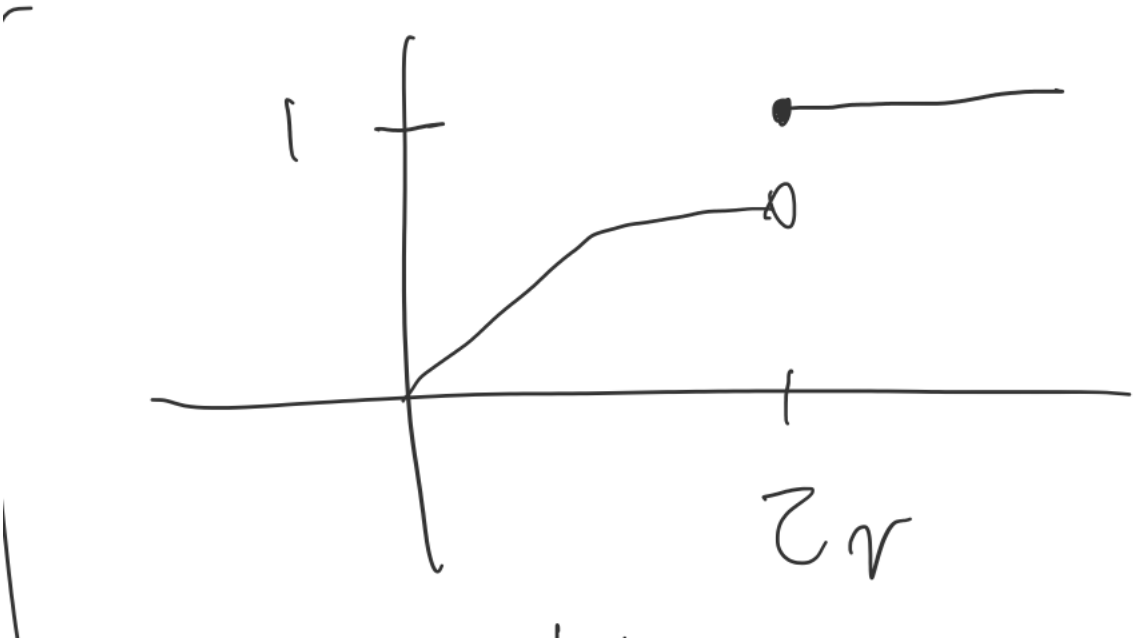
image5
Once I have the CDF for X, then I can compute anything I want about X, like its PMF, PDF, This CDF describes neither discrete nor continuous– it 's hybrid.

5 Lecture Note(11.2)

5.1 Hybrid Random Variable(Generalized PDF)

From last class working lifetime Ex.

$$F_x = \begin{cases} 0 & x \leq 0 \\ 1 - e^{-\frac{x}{\tau}} & 0 < x < \tau \\ 1 & x \geq \tau_r \end{cases}$$



This CDF is piece wise cont, with a finite number if jumps, This is a case of a "hybrid " Rand Var, since it has feature of both cont, and disr, RUs.

- -X has finite probability at jump discontinuities
- But elsewhere CDF is continuously non decreasing. Here X is continuously distributed

To compute quantities related to X, we explicitly write out a "generalized PDF". This takes the form

$$f_x(x) = f_{x,c}(x) + \sum_{x \in A(x)} a_j \delta(x - x_j)$$

Where the continuous part is

$$f_{x,c} = \frac{dF_x}{dx} \text{wherever the derivative exists.}$$

Don't worry about discontinuities

$$A(x) = \{x_1, x_2, \cdots, x_k\}$$

is the set of "atom", which the locations of jump. discontinuities
a_j is the magnitude of the jump of discontinuities.

$$\lim_{x\to x_j^+} F_x(x) - \lim_{x\to x_j^-} F_x(x) = F_x(x_j) - \lim_{x\to x_j^-} F(x) = Pr(X = x_j)$$

δ is "Dirac delta function which is simply a placeholder for locations of discontinuities"
For example

$$f_{x,c} = \begin{cases} 0 & x \le 0 \\ \frac{1}{\tau}e^{-\frac{x}{\tau}} & 0 < x < \tau_r, a(x) = \{\tau_r\} \\ 0 & x \ge \tau_r, a_1 = e^{-\frac{\tau_r}{\tau}} \end{cases}$$

Given a general pdf, we can compute expectation by a generalization of LUS.

$$\mathbb{E}[h(x)] = \int h(x) f_{x,c} dx + \sum_{x_j \in A(x)} a_j h(x_j)$$

For our example:

$$\mathbb{E}[x] = \int_0^{\tau_r} x \cdot (\frac{1}{\tau}e^{-\frac{x}{\tau}})dx + e^{-\frac{\tau_r}{\tau}}\tau_r$$

Some facts about computing problem with CDF

1. Since

$$\begin{aligned} 1 = \Pr(-\infty < X < \infty) &= \Pr(\{X > c\} \cup \{X \le c\}) \\ &= Pr(X > c) + Pr(X \le c) \\ &= Pr(X > c) + F(c) \\ \Pr(X > c) &= 1 - F(c)(cdf) \end{aligned}$$

2. For $b \ge a$

$$\begin{aligned} \Pr(x \le b) &= \Pr(X \le a) \cup \{X \in (a, b]\} \\ &= \Pr(X \le a) + \Pr(X \in (a, b]) \\ \Pr(x \in (a, b]) &= F(b) - F(a) \end{aligned}$$

Note: We cannot be sloppy about endpoint when working of CDF.(in discrete and hybrid). (Except continuous)

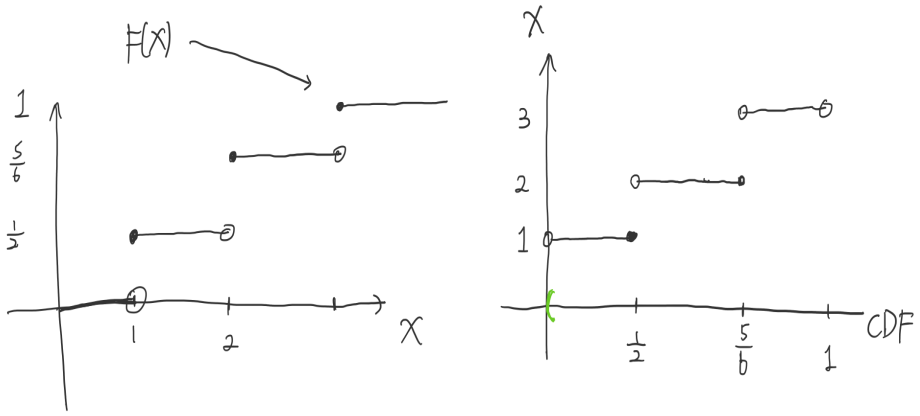
5.2 Quantile Function

We like a generalization of an inverse that can be applied to non-decreasing function, possibly with jumps, we'll denote this as F^{-1}, This is called the quantile function . formally.

$$F^{-1}(q) = \min\{y : q \le F(y) \text{ Traditionally we define domain of (0,1)}\}$$

Ex: Discrete Dist:

$R(X) = \{1, 2, 3\}, P_1 = \frac{1}{2}, P_2 = \frac{1}{3}, P_3 = \frac{1}{6}$



For continuous random variables, if CDF is 1:1 (strictly increasing), then F^{-1} is the traditional inverse of F
Ex: (Uniform dist) The pdf of unit dist of $[a, b]$ is

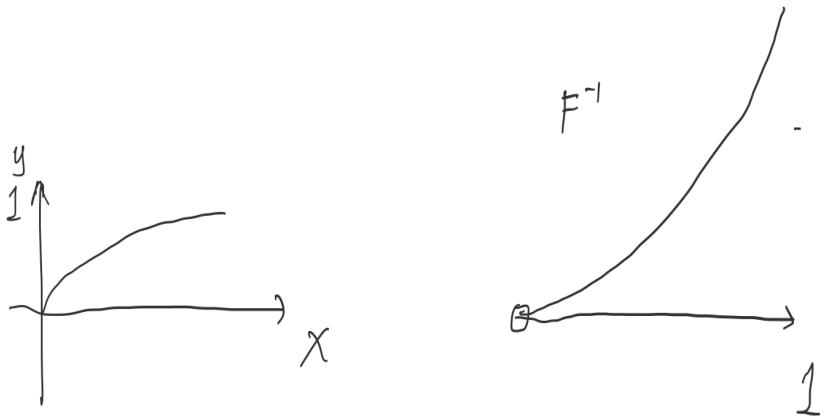
$$f(x) = \begin{cases} \frac{1}{b-a} & \text{on } x \in [a,b] \\ 0 & \text{otherwise} \end{cases} \longrightarrow F(x) = \begin{cases} 0 & \text{if } x \le a \\ \int_0^x \frac{1}{b-a} = \frac{x-a}{b-a} & \text{if } x \in [a,b] \\ 1 & \text{if } x \ge b \end{cases}$$

Then F(x) is strictly increasing over possible values of the Random variable To find $F^{-1}(x)$, solve $x = \frac{y-a}{b-a}$ for $y \to y = x(b-a) + a$
So $F^{-1}(q) = q(b-a) + a$
Ex:

$$F_X = \begin{cases} 1 - e^{-\frac{x}{\lambda}} & x \ge 0 \\ 0 & x < 0 \end{cases}$$

To compute quantile, find inverse

$$q = 1 - e^{-\frac{x}{\lambda}} \to 1 - q = e^{-\frac{x}{\lambda}} \to \log(1-q) = -\frac{x}{\lambda} \to = F^{-1}(q) q \in (0,1)$$

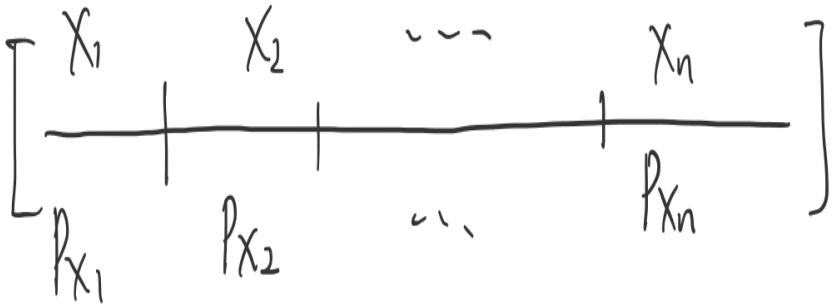


5.3 Inverse Transform Method(Quantiles)

Quantiles allow for us to simulate an arbitrary R.V X, provided that we have a uniform random generator [0,1] This is a widely used function in many programming language
Suppose $U \sim Unif(0,1)$ If I want to simulate random variable x, I first compute quatile F_x^{-1} and compose with U, giving $F_X^{-1}(U)$
We will show

$$X \sim F_x^{-1}(U)$$

For instance, to sample from exponential random variable with mean λ, we sample $X \sim -\lambda \log(1-U)$, **this is called inverse transform method(hw 1 d)**
In discrete case , this is equivalent to the following
Divide [0,1] into bins of size P_X, the probs of the pmf.
The bin that U falls into is the value of X simulated



Why does this work? Sppose we have RVs X and

$$Z = F_X^{-1}(U), U \sim Unif(0,1)$$

We show that X and Z have the same CDF, so $F_X(x) = F_Z(x)$

$$\begin{aligned} F_Z(X) &= \Pr(Z \le x) = \Pr(F_x^{-1} \le x) \\ &= \Pr(U \le F_x(x)) \\ &= F_x(x) \text{Thus: } F_z(x) \qquad\qquad = F_x(x) \end{aligned}$$

5.4 Back to Poisson Point Process

Nice fact about exponential distribution
We suppose random variable T that is exponential distribution, then it satisfies the "memoriless property"

$$\Pr(T > u | T > s) = \Pr(T > u - s) \qquad \text{for } 0 < s < u$$

This is stating that if I waiting for an event that is exponential distributed, having waited for time x does not affect the future probabilities.
To see this, note

$$
\begin{aligned}
\Pr(T > u | T > s) &= \frac{\Pr(\{T > u\} \cap \{T > s\})}{\Pr(T > s)} \\
&= \frac{\Pr(T > u)}{\Pr(T > s)} \\
&= \frac{1 - \Pr(T \leq u)}{1 - \Pr(T \leq s)} \\
&= \frac{1 - F_T(u)}{1 - F_T(s)} \\
&= \frac{1 - (1 - e^{-\frac{u}{\tau}})}{1 - (1 - e^{-\frac{s}{\tau}})} \\
&= \frac{e^{-\frac{u}{\tau}}}{e^{-\frac{s}{\tau}}} \\
&= e^{-\frac{u-s}{\tau}} \\
&= 1 - F_T(u - s) \\
&= \Pr(T > u - s)
\end{aligned}
$$

Memory less: Geometric The exponential is the only continuous random variable for this property
For Possion process, since the intervals between jumps here exp. dist, then intervals have memory less prop.
Some example of when PPP how better used:

- Time to arrived in a queue
- chemical reaction
- Price shock and stock market (simplified model)

6 Lecture Note(11.6)

6.1 Concept for part2 of this course

- Discrete random
  -
- continuous random variable
  - how to formulate in terms of probability density function , and how to compute probabilities of events involving continuous random variables in terms of the PDF
  - when the following continuous random variable models are appropriate:
    - Continuous uniform
    - Exponential
      - Connection to PPP, it is continuous analogue of the geometric distribution
      - Memoryless property
    - Normal(Gaussian)
  - Compute properties of random variables, particularly means and variances, standard deviations
    - scaling properties of means and variances under linear
- Moment generating functions
  -
- Cumulative distribution function (CDF)
  - Definition and the following 3 uses
    -
- PPP:
  - When it is an appropriate model, how the random variables from the class so far are associated to it
  - possion distribution is memoriless, we will wait until next incident happen.

We will begin the last unit of the class which
A common quantity related to multiple random variables is their sum. Give ...

$$\mathbb{E}Y = \sum_{j=1}^{n} \mathbb{E}X_j$$

The derivation of this formula just uses the fact that expectation is a linear operation, and the linear operations of summing and averaging commute(interchange sums or interchange sum and integral). The direct use of this formula is straightforward, but using the formula in reverse is a nice trick for many calculations of expectation of a random variable.
That is, we will focus on how to use this formula to compute the mean of a random ....
That is, one way to compute the mean of the complicated random variable Y
Example: Binomial Distribution We are given a binomially distributed random variable X with n trials and success probability p. We've computed its mean through the moment

generating function, but we'll now show another way to derive the mean of X by expressing X as a sum of simpler random variables.

$$X = \sum_{j=1}^{n} X_j \text{ where } X_j \text{ is a Bernoulli random variable with success probability p}$$

X_j indicates by a value of 1 a success on trial j and by a value 0 a failure on trial j. For this reason, Bernoulli random variables such as X_j are often called indicator random variables.

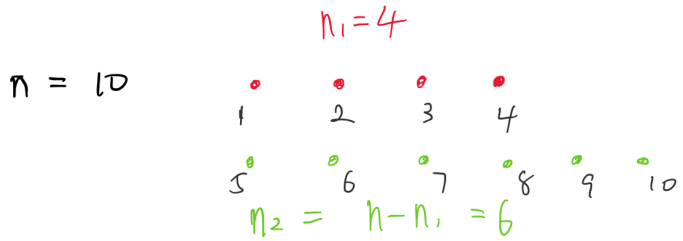
$$\mathbb{E}X = \mathbb{E}\sum_{j=1}^{n} X_j = \sum_{j=1}^{n} \mathbb{E}X_j$$

For a Bernoulli random variable with success probability $p, \mathbb{X}_j = p$

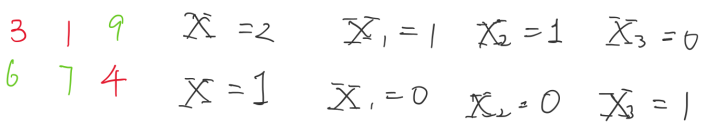
$$\mathbb{E}X = \sum_{j=1}^{n} p = np$$

6.2 Example: Hypergeomotric distribution

Suppose we are given a random variable X which has a hyper-geometric distribution based on dichotomous sampling without replacement of k selections from a population of size n, with one sub-population having n_1 members.



Ordered selections without replacement of size $k = 3$



Let's use the same strategy to express the number from the first sub-population as sum of Bernoulli/indicator random variables indicating what happened at each trial.

X_j will indicate whether trial j drew an item from the first sub-population

$$X = \sum_{j=1}^{k} X_j$$

(***Note the X_j affect each other but that doesn't matter for computing expectations.)

$$\mathbb{X} = \mathbb{E}\sum_{j=1}^{k} X_j = \sum_{j=1}^{k} \mathbb{E}X_j = \sum_{j=1}^{k} \Pr(X_j = 1)$$

What is $\Pr(X_j = 1)$? It's always $\Pr(X_j = 1) = \frac{n_1}{n}$ for all j. How show this?

- One could prove this by seeing it is true for $j = 1$, and the using law of total probability and induction to show it for $j > 1$
- Symmetry argument: Every trial in and of itself is selecting one out of the n items in the population with equal probability.
- Use classical probability with a sample space S of an ordered selection of k items without replacement from a population of size n. Let the event A be "The j^{th} item drawn is from the first sub-population"

$$\begin{aligned}\Pr(A) &= \frac{|A|}{|S|} \\ &= \frac{n_1 \times (n-1)_{k-1}}{(n)_k} \\ &= \frac{n_1(n-1)(n-2)...(n-(k-1))}{n(n-1)(n-2)...(n-(k-1))} \\ &= \frac{n_1}{n}\end{aligned}$$

So:

$$\begin{aligned}\mathbb{E}X &= \sum_{j=1}^{k} \Pr(X_j = 1) = \sum_{j=1}^{k} \frac{n_1}{n} \\ &= \frac{kn_1}{n}\end{aligned}$$

6.3 Example: Negative Binomial distribution

The negative binomial random variable X describe how many failures in a Bernoulli process with success probability p(and failure probability q=1-p) until the rth success. We can write $X = \sum_{j=1}^{r} X_j \text{where} X_j$ is the number of failures between the j-1st success and the jth success. (X_1 is the number of failures until the first success).

$$\mathbb{E}X = \mathbb{E}\sum_{j=1}^{r} X_j = \sum_{j=1}^{r} \mathbb{E}X_j$$

The random variable X_j are geometric with success probability p.
From mgf or whatever, we calculate $\mathbb{E}X_j = \frac{p}{q}$, so

$$\mathbb{E}X = \sum_{j=1}^{e} \frac{q}{p} = \frac{rq}{p}$$

We will get

6.4 Coupon collector experiment

Suppose we sample with replacement, repeatedly, from a population m elements, until we have selected k distinct objects from the population.

- See Coupon Collector Experiment app on "Random"

Let X be the random variable representing the number of trials requried to achieve this goal. This is similar, but not the same as, a negative binomial random variable, i.e., the number of failures before we obtain r successes in Bernoulli trials.
But the negative binomial distribution doesn't quite apply to the coupon collector problem because collecting is not the same as "success in a Bernoulli trial" Nonetheless, we can use the idea of decomposing the number of trials X required to collect k distinct coupon as:

$$X = \sum_{j=1}^{k} X_j$$

where X_j is the number of draws required after j-1 distinct coupons have been collected until k distinct coupons have been collected, i.e. until a novel coupon is selected.

$$X_j = 1 + Z_j$$

where Z_j is the number of "repeat" coupons drawn after $j-1$ distinct coupons have been collected until j distinct coupons have been collected.
Then Z_j would be a geometric distribution with success probability

$$p_j = \frac{m-(j-1)}{m}$$

Therefore $\mathbb{E}Z_j = \frac{q_j}{p_j} = \frac{1-p_j}{p_j} = \frac{1}{p_j} - 1 = \frac{m}{m-(j-1)} - 1$
So:$\mathbb{E}X = \mathbb{E}(\sum_{j=1}^{k} X_j) = \sum_{j=1}^{k} \mathbb{E}X_j = \sum_{j=1}^{k} 1 + \mathbb{E}Z_j = \sum_{j=1}^{k} 1 + \frac{m}{m-(j-1)} - 1 = \sum_{j=1}^{k} \frac{m}{m-(j-1)}$
If k is large(which also requires m to be large),then this sum looks like the Riemann sum of the integral.

$$\begin{aligned}
\approx \int_1^k \frac{m}{m-(j-1)} dx &= m \ln(m-(x-1))|_{x=1}^{k} \\
&= m(\ln(m-(k-1)) - \ln(m-(1-1))) \\
&= -m(\ln(m-k+1) - \ln m) \\
&= m \ln\left(\frac{m}{m-k+1}\right)
\end{aligned}$$

Comments: Coupon collector model can be useful way to think about analyzing some kinds of random sampling or computation algorithms. Break into the expected running time between checkpoints

6.5 Spore Model

Suppose a plant produces a chain of n spores, which then blows in the wind, and then spore chain will break ai any link between spores with probability p, independently of how the other links break

What is the expected number of chain fragments containing exactly l spores, where l can be sny number $1, \cdots n$? Call this random variable $X(l)$. It's the same as a fragment containing exactly $l-1$ consecutive unbroken links.
How can I break $X^{(l)}$ into a sum of random variables whose expectation is easy to calculate.

Write $X^{(l)} = \sum_{j=1}^{k} X_j^{(l)}$ where $X_j^{(l)}$ is a Bernoulli indicator random variable for the event that nodes $j, j+1, \cdots j+l-1$ from a fragment.

$$\Pr\left(X_j^{(l)} = 1\right) = (1-p)^{l-1}(\text{internal links unbroken}) \times p^2(\text{ links of ends of fragment broken})$$

using the independence of the fragments events to multiple probs.
For $2 \leq j \leq n-l+2$ (internal fragments)
$\Pr\left(X_j^{(l)} = 1\right) = (1-p)^{l-1} \times p^2$
For $j = 1, n-l+1$(end fragments, only one external link to break)

$$\Pr(X_j^{(n)} = 1) = (1-p)^{(l-1)}\text{the probability the whole chain remains a single fragment}$$

For Bernoulli/indicator random variables, $\mathbb{E}X_j^{(l)} = 1$
$\mathbb{E}X^{(l)} = \mathbb{E}\sum_{j=1}^{k} X_j^{(l)} = \sum_{j=1}^{k} \mathbb{E}X_j^{(l)} = \sum_{j=1}^{k} \mathbb{E}\Pr(X_j^{(l)} = 1)$
For $l < n$, two terms will be the endcases (j=1, j=n-l+1) and $n-l-1$ will be interior cases.
So using the above results:

$$\mathbb{E}X^{(l)} = 2(1-p)^{(l-1)} \times p + (n-l-1)(1-p)^{(l-1)} \times p^2$$

For the case

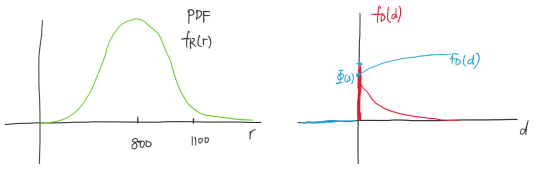
7 Lecture Note(11.9) Independent Random variables

7.1 Note on homework 4

1. Problem 2c: damage $D = g(R)$(damage as a function of rainfall)

$$D = c(\max(0, R - 1100))^{\frac{2}{3}} = g(R)$$

$$R \sim N(800, 100^2)$$



$$f_D(d) \neq f_R(g^{-1}(d))$$

2. Poisson process(Problem 2e)
   a.As we've defined in class, it is a continuum limit of Bernoulli process, which means that over different small time intervals, the probability for the incident to happen is the same small value and the occurrence of these incidents over different time intervals of the same size is independent.

   Figure
   There is a more general Poisson process that, to be more precise, is called an inhomogeneous Poisson process that only requires independence but allows the rate of occurrence to depend on time. But the time between incidents is then no longer exponentially distributed.

3. Poisson process is more generalized then Poisson distribution

7.2 Independent Random Variable

When working with the collection of random variables simultaneously, one generalyy needs to say something about how these random bariables are related to each other to do calculations, if we want anything other than the mean of the sum. The general way of expressing relationships between random variables is through joint probability distributions but we do not need to full technical framework when the random bariables are taken to be independent of each other.

A collection of random variables $\{X_1, X_2...x_n\}$ is said to be independent iff for any nice (Borel) subsets $B_j \subseteq S_j$ where S_j is the state space for X_j, we have:

$$\Pr(X_1 \in B_1, X_2 \in B_2, ...) = \Pr(X_1 \in B_1) \times \Pr(X_2 \in B_2) \times ... \times \Pr(X_n \in B_n)$$

In other words, the random variables$X_1, X_2, ..., X_n$ are independent precisely when the events $X_1 \in B_1, X_2 \in B_2, \cdots, X_n \in B_n$ are always independent.

A key result for independent random variables is that when we compute expectations involving multiple independent random variables, we have:

$$\mathbb{E}(X_1X_2...X_n) = \mathbb{E}X_1\mathbb{E}X_2....\mathbb{E}X_n$$

More generally, if we take functions $Y_j = g_j(X_j)$ where the g_j are deterministic, then the resuting random variables $Y_1, Y_2, ..., Y_n$ can also be shown to be independent of each other.

Putting these two facts together:

$$\mathbb{E}(g_1(X_1)g_2(X_2) \times ... \times g_n(X_n)) = \mathbb{E}g_1(X_1) \times \mathbb{E}g_2(X_2) \times ... \times \mathbb{E}g_n(X_n)$$

For deterministic functions $g_1, g_2, ..., g_n$
The results in red are only valid for independent random variables $X_1, ..., X_n$
Examples of collections of independent random variables:

- The amount of flood damage in year 1 and the amount of damage in year 2 if the rainfall in the two years is independent.
- The number of incidents in a Poisson process over non-overlapping intervals (# car incidents this week, # car accident next Friday between 4-5 PM)
- Sampling from a population with replacement
- The outcomes of different Bernoulli trials
- The times between successive incidents in a Poisson process

Not independent random variables:

- The amount of rainfall in a given year and the amount of flood damage in a given year
- # lizards born in a given year and the number of lizards caught in that year.
- The number of incidents in a Poisson process over overlapping intervals (# car accidents this week, # car accidents this month)
- Sampling from a population without replacement means the items selected at different draws are not independent(for example, particles binding to open slots)
  - only one particle can detach on same slot.
  - one slot can combine multiple particles

A very common operation on independent random variables is to add them up, possible with some deterministic weights:

$$Y = b + \sum_{j=1}^{n} c_j X_j \text{ with } X_j\text{independent random variables, and deterministic constants.}$$

Such sum arise naturally in statistics, as well as stochastic dynamics. We already know how to compute the mean of Y:

$$\mathbb{E}Y = b + \sum_{j=1}^{n} c_j \mathbb{E}X_j$$

true without any assumptions on the relationships of the X_j with each other

$$VarY = \sum_{j=1}^{n} c_j^2 Var(X_j) \text{ provided that } X_j \text{ are independent.}$$

To see why this is true, let's first do a subcalculation:
We already know from before that

(1) $Var(c_j X_j) = c_j^2 Var(X_j)$.

(2) $Var(b + Z) = Var(Z)$

(3) The above result will follow once we establish that

$Var(X_1 + X_2 + ... + X_n) = Var(X_1) + Var(X_2) + ... + Var(X_n)$ if $X_1, X_2, ...X_n$ independent

Because then:

$$Var(Y) = Var\left(b + \sum_{j=1}^{n} c_j X_j\right) \xrightarrow{(2)} Var\left(\sum_{j=1}^{n} c_j x_j\right) \xrightarrow{(3)} \sum_{j=1}^{n} Var(c_j x_j) \xrightarrow{(1)} \sum_{j=1}^{n} c_j^2 Var(x_j)$$

Just need to proof 3 that:

$$Var(X_1 + ... + X_n) = Var(X_1) + Var(X_2) + ... + Var(X_n)$$

It will follow by induction if we prove it for $n = 2$

$$\begin{aligned}
Var(X_1 + X_2) &= \mathbb{E}((X_1 + X_2) - \mathbb{E}(X_1 + X_2))^2 \\
&= \mathbb{E}((X_1 + X_2) - (\mathbb{E}X_1 + \mathbb{E}X_2))^2 \quad \text{Linearity of expectation} \\
&= \mathbb{E}((X_1 - \mathbb{E}X_1) + (X_2 - \mathbb{E}X_2))^2 \quad \text{regrouping} \\
&= \mathbb{E}((X_1 - \mathbb{E}X_1)^2 + 2(X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2) + (X_2 - \mathbb{E}X_2)^2) \\
&= \mathbb{E}(X_1 - \mathbb{E}X_1)^2 + 2\mathbb{E}[(X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2)] + \mathbb{E}(X_2 - \mathbb{E}X_2)^2 \\
&= Var(X_1) + 2\mathbb{E}[(X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2)] + Var(X_2)
\end{aligned}$$

We can seem $\mathbb{E}[(X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2)] \approx \mathbb{E}g_1(X_1)g_2(X_2)$. So we can use the result that says for deterministic g_1, g_2 and independent X_1, X_2

$$\begin{aligned}
\mathbb{E}[g_1(X_1)g_2(X_2)] &= \mathbb{E}g_1(X_2) \times \mathbb{E}g_2(X_2) \\
&= Var(X_1) + 2\mathbb{E}[(X_1 - \mathbb{E}X_1)(X_2 - \mathbb{E}X_2)] + Var(X_2) \\
E[(X_1 - \mathbb{E}X_1) \text{ looks like } \mathbb{E}(X_1 - a) &= \mathbb{E}_1 - a \\
&= Var(X_1) + 2(\mathbb{E}X_1 - \mathbb{E}X_1)(\mathbb{E}X_2 - \mathbb{E}X_2)] + Var(X_2)
\end{aligned}$$

7.3 Sums of independent, identically Distributed Random Variables

A collection of random variables $X_1, X_2, ..., X_n$ is said to be identically distributed if they all have the same probability distribution (same PDF or PMF or CDF).

- for example, the indicator variables for sampling from the list of subpopulation are identically distributed both with and without replacement.

- the number of trials needed to get one more coupon in the coupon collector problem are not identically distributed (even though they are independent)

We say a collection of random variables $\{X_1, X_2, ...X_n\}$ are said to be iid if they are independent and identical distributed

- most common occurrence of iid random variables in practice is through repeated experiments and/or simulations of the same process on different realizations.
- As "white noise" source

Sum of iid random variables have important simple expressions:

$$S_n = \sum_{j=1}^{n} X_j \text{ with } X_j \text{ iid with } \mu = \mathbb{E}X_j \text{ and } \sigma^2 = Var(X_j)$$

$$\mathbb{E}S_n = \sum_{j=1}^{n} \mathbb{E}X_j = \sum_{j=1}^{n} \mu$$

$$\mathbb{E}S_n = n\mu$$

Similarly,

$$VarS_n = \sum_{j=1}^{n} VarX_j = \sum_{j=1}^{n} \sigma^2$$

$$Var(S_n) = n\sigma^2 \text{ where } \sigma^2 = VarX_j$$

$$\text{Also } \sigma_{S_n} = \sqrt{n}\sigma$$

This has an important consequence for the sample of a collection of iid random variables

$$\hat{\mu} = \frac{S_n}{n} = \frac{1}{n}\sum_{j=1}^{n} X_j$$

This is of course how we try to estimate the mean of $\mathbb{E}X$ from iif samples $\{X_j\}^n$, drawn from the probability distribution for X, Notice the sample mean $\hat{\mu_n}$ is random! What are its statistical properties?

$$\mathbb{E}\hat{\mu} = \mathbb{E}\left(\frac{S_n}{n}\right) = \frac{1}{n}\mathbb{E}S_n = \frac{1}{n}(n\mu) = \mu$$

This is, in statistics, the property of an unbiased estimator.

- randomly select samples from bins. we will get approx ratio of bin

What about the fluctuations of the sample mean?

$$\sigma_{\hat{\mu}} = \frac{1}{n}\sigma_{S_n} = \frac{1}{n}(n\mu) = \mu$$

Standard deviation of the sample mean is:

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$$

This is what's known as sampling error in Monte Carlo simulations, and results from taking only a finite number of samples which incompletely samples the randomness. The sampling error will decrease according to the inverse squire root of the amount of effort (# simulations or # experiments or # observatons). This relationship between sampling error and effort is fundamental and one of the biggest disadvantages of Monte Carlo simulations. Most deterministic computational methods have much better payoff for accuraxcy w.r.t effort. However , deterministic method have a huge overhead cost to implement or even run at low "low effort" and/or produce garbage with low effort.

1. Some people advocated using "quasi-Monte Carlo" methods to improve somewhat on the $\frac{1}{\sqrt{n}}$ improvement with effort

8 Lecture note 11/16

Sums of iid random variable

$$\mathbb{E}\hat{\mu}$$

This is, instat
Standard deviation of the sample mean

$$\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$$

This is what's known as sampling error in Monte Carlo simulations, and results from taking only a finite number of samples, which incompletely samples the randomness. The sampling error will decrease according to the inverse square root of the amount of effort (# simulations or # observations). This relationship between sampling error and effort is fundamental and one of the biggest disadvantages of Monte Carlo simulations. Most deterministic computational methods have much better payoff for accuracy w.r.t. effort. However, deterministic methods often have a huge overhead cost to implement or even run at "low effort" and/ or produce garbage with low effort.

- Some people advocated using "quasi=Monte Carlo" methods to improve somewhat on the $\frac{1}{\sqrt{n}}$ improvement with effort. figure

By the way, you can estimate the standard deviation of a probabilistic distribution from random sample of size n through the following formula sample standard deviation:

$$\hat{\sigma} = \sqrt{\frac{\sum_{j=1}^{n}(X_j - \hat{\mu})^2}{n-1}}$$

n-1: minus the degree of freedom.
Look at the applet called "Sample Mean Experiment" The other component of the sampling error is the factor σ which is just a measure of how fundamentally random a realization is.

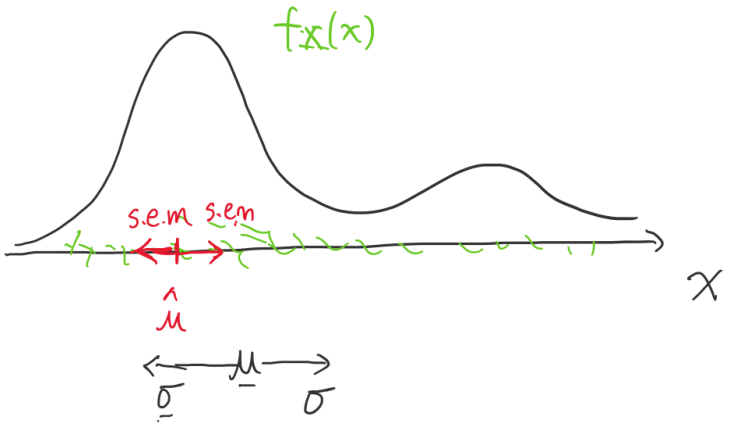
- Variance reduction methods try to design the realizations to reduce σ

At more elementary level, we at least have that $\sigma_{\hat{\mu}} \to 0$ as $n \to \infty$ The property of no bias and the sampling error converging to 0 with effort is what's known as the sample mean estimator being consistent.
By the way, how would you estimate the error of your sample mean?
The true standard deviation of the sample mean is $\sigma_{\hat{\mu}} = \frac{\sigma}{\sqrt{n}}$ but I don't know σ. So rather let's use the guess for σ from the sample standard deviation, and this gives us the standard error of the mean (s.e.m) which is the estimated standard deviation of the sample mean:

$$s.e.m" = "\hat{\sigma}_{\hat{\mu}} = \frac{\hat{\sigma}}{\sqrt{n}}$$

So a good practice in reporting the results for Monte Carlo simulations of some model or calculation, if you did n simulations or experiments is to report your estimate for the computed quantity as : $\hat{\mu} \pm \hat{\sigma}_{\hat{\mu}}$ Or if you are plotting the results, plot the sample mean as your best guess anf make the error bars some multiple of $\sigma_{\hat{\mu}}$

Some multiple? Depends what field you are in. In applied



$$\lim_{n\to\infty} \hat{\mu} = \mu$$

This is just an expression of the Law of Large Numbers and is the rigorous basis of frequentist thinking regarding probability.

Moreover, sums of iid random variables will approach a normal (Gaussian) distribution as $n \to \infty$; this is the Central Limit Theorem

$$\lim_{n\to\infty} \frac{S_n = n\mu}{\sigma\sqrt{n}} \sim N(0,1) \text{ as } n \to \infty$$

Often in random algorithms in computer science, one wants to make provable statements about performance. And these often take the form that the answer is correct with some high probability, or the answer has a sort of specific level of accuracy with some high probability. The high probability is expressed in terms of parameters of the algorithm and. or the effort used(i.e., the number of items sampled). How are these statements derived? One fundamental tools is the Bernstein inequality. Bernstein inequality gives a bound for the probability that a sum of iid random variables differs by some amount from its mean.

So far we've presented formulas for computing means and standard deviations of sums of independent random variables. What about the probability distribution of $Y = \sum_{j=1}^{n} X_j$ where X_j are independent? This question can not be answered quite so easily as the mean and standard deviation for Y
Consider first the simplest case where X_j are discrete, and $n = 2$,
We are given the pmfs of X_1, X_2, maybe they're different.

$$p_x^{(X^1)} = \Pr(X_1 = x)$$
$$p_x^{(X^2)} = \Pr(X_2 = x)$$

Clearly $Y = X_1 + X_2$ should also be discrete. What is its pmf?

$$p_y^{(Y)} = \Pr(Y = y) = \Pr(X_1 + X_2 = y)$$

We chould do this by a brute force summing up of all the elementarty events$(X_1 = x_1, X_2 = x_2)$ that belong to the event $A = \{X_1 + X_2 = y\}$
Or by law of total probability, partitioning on the values of X_2:

$$\begin{aligned}\Pr(X_1 + X_2 = y) &= \sum_{x_2 \in R(X_2)} \Pr(X_1 + X_2 = y | X_2 = x_2) \Pr(X_2 = x_2) \\ &= \sum_{x_2 \in R(X_2)} \Pr(X_1 + x_2 = y | X_2 = x_2) p_{x_2}^{(X_2)} \\ &= \sum_{x_2 \in R(X_2)} \Pr(X_1 = y - x_2 | X_2 = x) p_x^{(X_2)}\end{aligned}$$

But since $X_1 m X_2$ are independent, the events $X_1 = y - x, X_2 = x$ are independent

$$\begin{aligned}\Pr(X_1 + X_2 = y) &= \sum_{x_2 \in R(X_2)} \Pr(X_1 = y - x_2) p_x^{(X_2)} \\ &= \sum_{x_2 \in R(X_2)} p_{y-x}^{(X_1)} p_x^{(X_2)} \text{ It is a convolution of the probability mass functions of } X_1, X_2\end{aligned}$$

What if you add up n independent random variables. If you repeat the above argument recursively, you would find that the sum of n independent random variables is the (n-1) That is, sum over all values in $n-1$ nested operations. But there is an efficient way to compute convolutions: Fourier transform, Laplace transform convert convolutions to multiplication operations.

Moment generating functions behave like Laplace transforms for random variables. And because we saw that the probability distribution of sums of independent rvs involve convolution, mgf will be a good way to treat independent random variable

A good alternative procedure for working with sums of independent random variables is to use moment generating function. Moment generating functions(mgfs) are essentially Laplace transforms of the probability distribution, and so the reason mgfs are useful for sums of independent random variables is the same reason that Laplace/Fourier transforms simplify convolutions.
Moment generating function:

$$M_Y(s) \mathbb{E} e^{sY}$$

Let's try to calculate it in terms of the known random variables X_1, X_2 :

$$\begin{aligned}M_Y(s) &= \mathbb{E} e^{sY} \\ &= \mathbb{E} s^{s(X_1 + X_2)} \\ &= \mathbb{E}[e^{sX_1} e^{sX_2}] \\ &= \mathbb{E} e^{sX_1} \mathbb{E} e^{sX_2} \\ &= M_{x_1}(s) M_{X_2}()s\end{aligned}$$

9 Note 11.20 Happy thanksgiving

The mgf of a sum of independent randomvariables is therefore easily related
Illustration: Suppose X_1 and X_2 are two independent poisson random bariables, with means λ_1, λ_2 what is probability distribution for $Y = X_1 + X_2$

$$\begin{aligned}
\Pr(Y = y) &= \sum_{x \in R(X_2)} p_{y-x}^{(X_1)} p_x^{(X_2)} \\
p_x^{X_1} &= \frac{e^{-\lambda}\lambda_1^x}{x!}; p_x^{X_2} = \frac{e^{-\lambda}\lambda_2^x}{x!} \text{ for x=0,1,2,} \cdots \\
\Pr(Y = y) &= \sum_{x=0}^{\infty} p_{y-x}^{X_1} p_x^{(X_2)} \\
&= \sum_{x=0}^{y} \frac{e^{-\lambda}\lambda_1^{y-x}}{(y-x)!} \frac{e^{-\lambda_2}\lambda_2^x}{x!} \\
&= \frac{e^{-\lambda_1-\lambda_2}}{y!} \sum_{x=0}^{y} \frac{y!}{(y-x)!} \frac{1}{x!} \lambda_1^{y-x} \lambda_2^x \\
&= \frac{e^{-\lambda_1-\lambda_2}}{y!} \sum_{x=0}^{y} \binom{y}{x} \lambda_1^{y-x} \lambda_2^x \\
&= \frac{e^{-\lambda_1-\lambda_2}}{y!} (\lambda_1 + \lambda_2)^y \text{ by binomial theorem, for } y = 0, 1, 2, ...
\end{aligned}$$

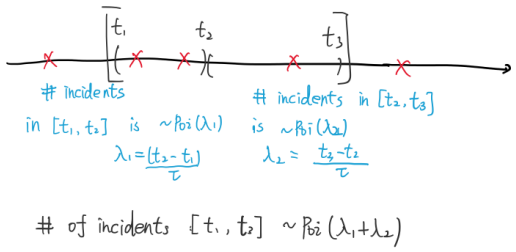
So we see from the pmf of $Y = X_1 + X_2$ that Y is Poisson rv with mean $\lambda_1 + \lambda_2$.
Let's re-derive this result using mgfs.
We derived the mgf of a Poisson random variable in a previous lecture:

$$M_{X_1}(s) = e^{\lambda_1(e^s-1)}, M_{X_2}(s) = e^{\lambda_2(e^s-1)}$$

$$M_Y(s) = M_{X_1}(s) M_{X_2}(s) = e^{\lambda_1(e^s-1)} e^{\lambda_2(e^s-1)} = e^{(\lambda_1+\lambda_2)(e^s-1)}$$

This is the same as the mgf for a Poisson rv with mean $\lambda_1+\lambda_2$.And mgfs uniquely characterize the probability distribution. Therefore this shows that Y is a Poisson random variables with mean $\lambda_1 + \lambda_2$
This direct calculation agree with the intuition we should have about Poisson random variables, particularly in connection with the Poisson process
(figure)
If we are adding together n independent random variables: $Y = \sum_{j=1}^{n} X_j$the induction:

$$M_Y(s) = \Pi_{j=1}^{n} M_{X_j}(s)$$



Applying this to some of our common probability distributions, we find the following relationships

- adding independent Poisson random variables gives a Poisson random variable
- adding independent Gaussian/normal random variables gives a Gaussian/normal random variable
- adding independent binomial random variables together, if they have the same success probability, again gives a binomial random variable
- Adding independent geometric and/ or negative binomial random variables with the same success probabilities gives negative binomial random variable.

More generally computing the probability distribution for a sum of independent random variables via mgfs has cost scaling linearly with n, while the direct calculation with repeated convolutions would have cost scaling as m^n where m is a large value (the range of the probability distributions.)
Also the Central limit theorem is essential proved by using mgfs(usually characteristic functions) and Taylor expansion about small s(k)m and what one shows is that multiplying the mgfs of n iid random variables together produces and mgf that converges to the mgs of a Gaussian as $n \to \infty$. The technical requirements have to do with how the mgfs behave near $s = 0(k = 0)$, which requires that the random variables have probability distribution

These statements are all consistent because Poisson distribution, binomial distribution, and negative binomial distribution with large means can all be shown to be approximately Gaussian.

- check this visually with Special Distribution Simulator app

9.1 Dependent Multiple Random Variables

If random bariables are not independent, then we must develop a technical framework for describing their relationship. Just as for individual random variables, we will first develop a comprehensive description, and then simpler incomplete but useful descriptions of the relationship.
The comprehensive description a collection of random variables $\{X_1, X_2 \cdots X_n\}$ is by their joint probability distribution. We'll begin by focusing on discrete random variables, in which case the joint probability distribution is described by a joint mass function (PMF)

$$p_{x_1,x_2,...,x_n} = \Pr(X_1 = x_1, X_2 = x_2, ..., X_n = x_n)$$

for all possible $x_j \in R(X_j)$ The marginal probability distribution for any one of these random variables is given by:

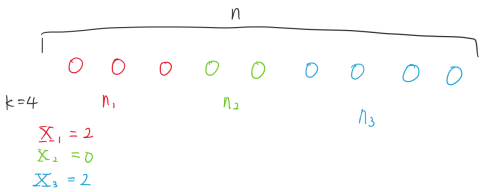
$$p_x^{(X_j)} = \Pr(X_j = x)$$

This is just viewing the random variableX_j on its own, without regard to the other related random variables.
We know that for independent random variables, the joint probability distribution is the product of the marinal probability distributions

$$p_{x_1,x_2,\ldots,x_n} \prod_{j=1}^{n} p_{x_j}^{X_j}$$

Joint PMF is product of marginal PMF but this is not true for general collections of random variables.
Examples where we have a collection of random variables which are not independent: Sampling from a population with r sub-populations.

Suppose we have a population of n items, which can be decomposed into r subpopulations, with n_j items in subpopulation j, and $n_1 + n_2 + \cdots + n_r = n$



Sample k items from this population without replacement, and define X_j to be the number of items selected from the jth sub-population, Are the random variable $\{X_j\}_{j=1}^r$ independent?
Can't be because for example knowing $X_3 = 4$ in the above sample would make $X_1 = 1$ impossible, but knowing $X_3 = 3$ would make $X_1 = 1$ possible with positive possibility. So the pmf of X_1 depends on the value of X_3 so those random variables can't be independent.
Let X_j denote the number of items selected from the jth subpopulation, for $j = 1, ..., r$
The joint probability distribution will generalize from the hyper-geometric distribution using the counting principle

$$p_{x_1,x_2,\ldots,x_r} = \begin{cases} \frac{\binom{n_1}{x_1}\binom{n_2}{x_2}\times\cdots\times\binom{n_r}{x_r}}{\binom{n}{x}} & \text{for } x_1 + x_2 + ... + x_r = k 0 \leq x_i \leq n_i \\ 0 & \text{Otherwise} \end{cases}$$

This is called the extended hyper-geometric distribution.remember to look the range of the random variable
Suppose we just want to examine whether or not the random variables $X_1, X_2, \cdots X_r$ are independent from mathematics alone?
It sort of looks like a product of a function of x_1 times a function of x_2 times a function of x_3 .. but we concluded that these random variable can't be independent. But the constraints$x_1 + x_2 + ... + x_r = k$ ties the random variables together and makes the joint pmf not the product of marginals.
To see this more precisely, let's look at the marginal distribution for X_j

$$p_x^{X_j} = \Pr(X_j = x) = \frac{\binom{n_j}{x}\binom{n-n_j}{k-x}}{\binom{n}{k}} \text{ for } 0 \leq x \leq k, n_j$$

This is just the hyper-geometric distribution for sampling without replacement from a population with two subpopulation: the jth subpopulation and everything else.

It is now clear that the joint PMF is not the product of the marginal PMF.That proves that they are not independent.
Let's now consider sampling from the population with r sub-population , but now with replacement.
Joint pmf is:

$$p_{x_1,x_2\cdots x_r} = \binom{k}{x_1, x_2, ..., x_r} \frac{n_1^{x_1} \cdot n_2^{x_2} \cdots n_r^{x_r}}{n^k} \text{ for } 0 \leq x_i \leq k, x_1 + x_2 + ... + x_r = k$$

This is the multi-nomial distribution, which can be written in more standard form by writing the probability to choose from subpopulation i as $p_i = \frac{n_i}{n}$

$$\begin{aligned} p_{x_1,x_2\cdots x_r} &= \binom{k}{x_1, x_2, ..., x_r} p_1^{x_1} p_2^{x_2} \times ... \times p_r^{x_r} \text{ for } 0 \leq x_i \leq k, x_1 + x_2 + ... + x_r = k \\ &= \frac{k!}{x_1! x_2! \cdots x_r!} p_1^{x_1} p_2^{x_2} \times ... \times p_r^{x_r} \end{aligned}$$

Again we have a constraint that ties together the values of the different random variables, so this joint pmf does not look like it describes independent random variables, even though the actual value of $p_{x_1,x_2\cdots x_r}$ look like they factor
What is the margin distribution for X_j when I sample with replacement? Again we just think about the jth subpopulation and "everything else" and so the probability distribution of the number drawn from the jth subpopulation is equivalent to drawing from a population with two sub-populations, which has a binomial distribution.

$$\begin{aligned} p_x^{(X_j)} &= \Pr(X = x) \\ &= \binom{k}{x} p_j^x (1 - p_j)^{k-x} \text{ for } x = 0, 1, 2 \cdots k \end{aligned}$$

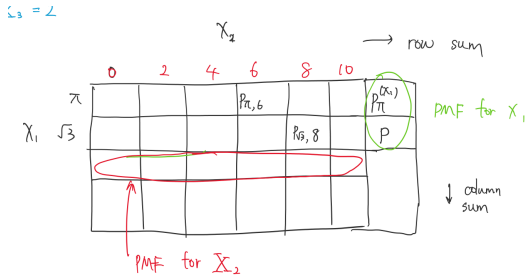
Joint PMF is not the product of the marginal PMFS
Above we computed the marginal PMFs from thinking about the model.But that shortcut won't always be possible, and you should also know the systematic brute force way to get marginal PMFs from joint PMFs.
If I know the joint PMF $p_{x_1,x_2,...,x_n}$ of $X_1, X_2, ...X_n$
Then I can derive the marginal PMF of any of these random variables by:

$$p_x^{X_j} = \Pr(X_j = x) = \Pr\left(\bigcup_{x \in \mathbb{R}(X_i), i \neq j} (X_1 = x_1, X_2 = x_2, ...X_j = x_j, ...X_n = x_n) \right) \text{ disjoint union}$$

$$= \sum_{x_i \in R(X_i),\ i \neq j} \Pr(\mathbb{X}_1 = x_1, \mathbb{X}_2 = x_2, ...X_j = x_j, ...X_n = x_n)$$

$$p_x^{X_j} = \sum_{x_i \in R(X_i), i \neq j} p_{x_1,x_2,...,x,...,x_n} \text{ x is in the jth place}$$

If you think about this for $n = 2$, it can be seen graphically and explains why marginal PMFs are called marginal.



Let's see how this systematic computation of marginal PMF from joint PMF works for the example above with sampling with replacement from a population with r sub-populations. Start with joint PMF:

$$p_{x_1,x_2\cdots x_r} = \binom{k}{x_1, x_2, ..., x_r} p_1^{x_1} p_2^{x_2} \times ... \times p_r^{x_r} \text{ for } 0 \leq x_i \leq k, x_1 + x_2 + ... + x_r = k$$

Let's get the marginal PMF for X_1 from this without thinking about the model anymore.

$$\begin{aligned} p_X^{(X_1)} &= \Pr(X = x) \\ &= \sum_{0 \leq x_1, x_2 \cdots x_r \leq k, x_1 + x_2 + \cdots x_r = k} \binom{k}{X, x_2, ..., x_r} p_1^{x_1} \\ &= \sum \frac{k!}{X! x_2! \cdots x_r!} p_1^X p_2^{x_2} \cdots p_r^{x_r} \\ &= \frac{p_1^x}{x!} \sum_{....} \frac{k!}{x_2! x_3! ... x_r!} p_2^{x_2} \cdots p_r^{x_r} \end{aligned}$$

10 Lecture Note(11.27) Conditional Probability Distributions

10.1 Conditional Probability Distributions

The joint probability distribution, though it has all needed information about the random variables, can be awkward to work with. A related construction which is particularly useful for calculations involving multiple random variables are conditional probability distributions. For a general event A, we can define:

$$p_x^{(X|A)} \equiv \Pr(X = x|A) \frac{\Pr(\{X = x\} \cap A)}{\Pr(A)}$$

A particular important case is to condition on the value of another random variable

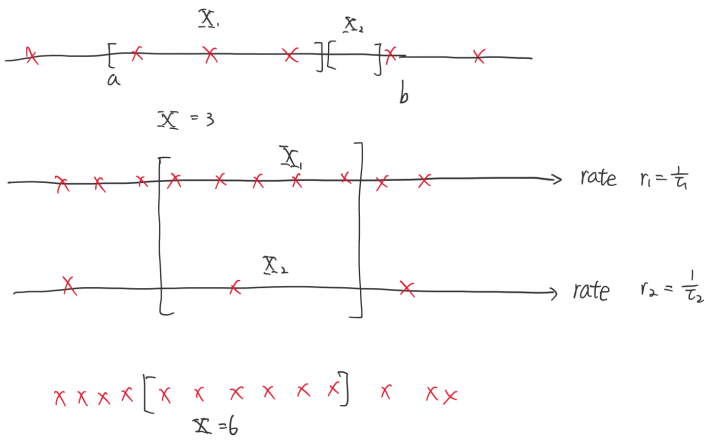
$$\begin{aligned} p_{x|y}^{(X|Y)} &\equiv \Pr(X = x|Y = y) \\ &= \frac{\Pr(\{X = x, Y = y\}}{\Pr(Y = y)} (\text{joint/marginal}) \\ &= \frac{p_{x,y}}{p_y^{(Y)}} \end{aligned}$$

This is known as the conditional probability distribution (or for discrete rvs, the conditional PMF) of X given Y.It's just the ratio of their joint probability distribution to the marginal distribution of the conditioned variable Y.$\cdots$

Example of computing a conditional probability distribution:

Suppose we have two independent Poisson random variables X_1, X_2, with means λ_1, λ_2. What is the condition probability distribution of X_1 given a certain of their sum $Y = X_1 + X_2 = y$? Two application of such calculation:

1. Consider a Poisson process, take an interval [a,b] and write it as the union of two sub-intervals. Let Y be the number of Poisson points (successes) in the interval [a,b], and X_1, X_2 be the number of successes (Poisson points) in each of the two sub-intervals
2. Or we could imagine having two independent Poisson processes, and looking at the total number of successes from both over some time interval. Given that observation, how many came from the first Poisson process.



Recall from our previous discussion about the sums of independent Poisson random variable that Y will also be a Poisson random variable with mean $\lambda_1 + \lambda_2$
Calculate the conditional PMF of X_1 given Y(can't be independent since)

$$\Pr(A, B) = \Pr(A \cap B)$$

$$\begin{aligned}
p_{x_1|y}^{(X_1|Y)} &\equiv \Pr(X_1 = x_1 | Y = y) \\
&= \frac{\Pr(\{X_1 = x_1, Y = y\}}{\Pr(Y = y)} \\
\Pr(\{X_1 = x_1, Y = y\} &\xrightarrow{\text{re-express joint}} \Pr(X_1 = x, X_1 + X_2 = y) \\
&\xrightarrow{\text{calculation in terms of independent RVS}} \Pr(X_1 = x, x + X_2 = y) \text{comma means and} \\
&= \Pr(X_1 = x, X_2 = y - x) \\
&\xrightarrow{\text{independence}} \Pr(X_1 = x) \Pr(X_2 = y - x) \\
&= p_x^{(X_1)} p_{y-x}^{X_2} \\
\Pr(Y = y) &= p_y^{(Y)} \\
&= \frac{e^{-\lambda_1 - \lambda_2} (\lambda_1 + \lambda_2)^y}{y!} \text{for } y = 0, 1, 2 \cdots
\end{aligned}$$

Because$Y \sim Poi(\lambda_1 + \lambda_2)$

$$p_{x_i}^{(X_i)} = \frac{e^{-\lambda_i} (\lambda_i)_i^x}{x_i!} \text{for } x_i = 0, 1, 2 \cdots$$

Therefore

$$\begin{aligned}
p_{x_1|y}^{(X_1|Y)} &\equiv \Pr(X_1 = x_1 | Y = y) \\
&= \frac{\Pr(\{X_1 = x_1, Y = y\}}{\Pr(Y = y)} \\
&= \frac{p_x^{(X_1)} p_{y-x}^{X_2}}{p_y^{(Y)}} \\
&= \frac{\frac{e^{-\lambda_1} \lambda_1^x}{x!} \frac{e^{-\lambda_2} \lambda_2^{y-x}}{(y-x)!}}{\frac{e^{-\lambda_1 - \lambda_2} (\lambda_1 + \lambda_2)^y}{y!}} \\
&= \frac{\lambda_1^x \lambda_2^{y-x}}{(\lambda_1 + \lambda_2)^y} \frac{y!}{x!(y-x)!} \\
&= \binom{y}{x} \left(\frac{\lambda_1}{\lambda_1 + \lambda_2} \right)^x \left(\frac{\lambda_2}{\lambda_1 + \lambda_2} \right)^{y-x} \quad \text{for } x = 0, 1, 2, ..., y
\end{aligned}$$

Therefore,$X_1|Y \sim Bin(Y, \frac{\lambda_1}{\lambda_1 + \lambda_2})$ Therefore, given $Y = X_1 + X_2 = y$, then X_1 will be distributed binomially with y trials and success probability $p = \frac{\lambda_1}{\lambda_1 + \lambda_2}$
Tips In computing $\Pr(A, B)$ can use A to simplify B
Besides computing conditional probabilities on demand(i.e., when the problem is asking you to calculate a property of a random variable given some partial)

The key formula is the law of total probability, which in terms of PMFs looks like:

$$p_x^{(X)} = \sum_{j=1}^{r} p_x^{(X|B_j)} \Pr(B_j)$$

When $\{B_j\}_{j=1}^r$ is a partition of the sample space.
Often we define a partition by the value some other random variable Y takes

$$p_x^X = \sum_{y \in S_y} p_{x|y}^{X|Y} p_y^{(Y)}$$

Example: Suppose X is the number of successes in N Bernoulli trials with success probability p, where N is itself a random variable which is described by a Poisson distribution with mean λ. What is the (marginal) probability distribution for X ?

Interpretation: Now imagine that we have a Poisson process with rate r and each Poisson point is designated to be "special" with probability p, independently of the designation of every other point. Then after time t, N would describe the total number of Poisson points with mean $\lambda = rt$, and X would denote the number of those point that are special.
figure

This procedure is called "thinning a Poisson process"

$$p_x^{(X)} = \Pr(X = x) = \sum_{n=0}^{\infty} p_{x|n}^{X|N} p_n^{(N)}$$

We did this because knowledge of N makes the properties of X much easier to calculate.

$$\begin{aligned}
p_n(N) &= \frac{e^{-\lambda}\lambda^n}{n!} \text{ for } n = 0, 1, 2 \cdots \\
p_{x|n}^{X|N} &= \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, 2, ..., n \\
p_x^X &= \sum_{n=x}^{\infty} \binom{n}{x} p^x (1-p)^{n-x} \frac{e^{-\lambda}\lambda^n}{n!} \\
&\text{n=x because } p_{x|n}^{X|N} = 0 \text{ unless } x \leq n \\
&= e^{-\lambda} \sum_{n=x}^{\infty} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \frac{\lambda^n}{n!} \\
&= \frac{e^{-\lambda} p^x}{x!} \sum_{n=x}^{\infty} \frac{1}{(n-x)!} (1-p)^{n-x} \frac{\lambda^n}{1}
\end{aligned}$$

Change dummy index to be $m = n - x$

$$= \frac{e^{-\lambda}p^x}{x!}\sum_{m=0}^{\infty}\frac{\lambda^{m+x}(1-p)^m}{(m)!}$$

$$= \frac{e^{-\lambda}(p\lambda)^x}{x!}\sum_{m=0}^{\infty}\frac{\lambda^{m}(1-p)^m}{(m)!}$$

$$= \frac{e^{-\lambda}(p\lambda)^x}{x!}e^{(1-p)\lambda}$$

$$p_x^X = e^{-p\lambda}\frac{(p\lambda)^x}{x!} \text{ for } x = 0, 1, 2, ...\text{which is a Poisson distribution with mean } p\lambda$$

The above calculations, together with verification of certain independence properties (that are fairly obvious) show that: Merging two independent Poisson with rates r_1 and r_2 gives a Poisson process with rate $r_1 + r_2$.
Thinning a Poisson process with rate r and thinning probability p gives a Poisson process with rate pr

10.2 Conditional Expectation

$$\mathbb{E}[X|A] = \mu_{(X|A)} \equiv \sum_{x\in S_x} xp_x^{(X|A)}$$

It is just the average with respect to the conditional probability distribution of X given A
In particular, if we want to condition on the value of another random variable:

$$\mathbb{E}[X|Y = y] =< X|Y = y >\equiv \sum_{x\in S_X} xp_{x|y}^{X|Y}$$

When we partition on the value of another random variable Y, we can write this as

$$\mathbb{E}X = \sum_{y\in S_Y}\mathbb{E}[X|Y = y]p_y^Y$$

or in the advanced notation:$\mathbb{E}X = \mathbb{E}(\mathbb{E}[X|Y])$.
Related to the law of total expectation, we have the following generalization

$$\mathbb{E}g(X) = \sum_{j=1}^{r}\mathbb{E}[g(X)|B_j]Pr(B_j)$$

$$\mathbb{E}g(X) = \sum_{y\in S_Y}\mathbb{E}[g(X)|Y = y]p_y^{(Y)}$$

This follows from the standard law of total expectation because $g(X) \equiv Z$ is just a random variable too.
Example: Recall the problem we just studied.
Suppose X is the number of successes in N Bernoulli trials with success probability p, where N is itself a random variable which is described by a Poisson distribution with mean λ. What is the (marginal) probability distribution for X ?

Let's try this time to compute the mgf of X

$$\begin{aligned}
M_X(s) &= \mathbb{E}e^{sX} \\
&= \sum_{n=0}^{\infty} \mathbb{E}[e^{sX}|N=n]p_n^{(N)} \\
p_n^{(N)} &= \frac{e^{-\lambda}\lambda^n}{n!} \text{ for } n = 0, 1, 2, ... \\
\mathbb{E}[e^{sX}|N=n] &= (pe^s + (1-p))^n \text{ which is the mgf for Bin(n,p)} \\
M_X(s) &= \sum_{n=0}^{\infty} (pe^s + (1-p))^n \frac{e^{-\lambda}\lambda^n}{n!} \\
&= e^{-\lambda} \sum_{n=0}^{\infty} (pe^s + (1-p))^n \frac{\lambda^n}{n!} \\
&= e^{-\lambda} \sum_{n=0}^{\infty} (\lambda(pe^s + (1-p)))^n \frac{1}{n!} \\
&= e^{-\lambda} e^{\lambda(pe^s + (1-p))} \\
&= e^{-\lambda + \lambda(pe^s + (1-p))} \\
&= e^{\lambda pe^s - \lambda p} \\
&= e^{\lambda p(e^s - 1)}
\end{aligned}$$

This is the mgf of a Poisson distribution with mean λp. Mgfs are unique descriptors of random variables, therefore $X \sim Poi(\lambda p)$
Why did the mgf work so well here? Poisson processes have independence properties and mgfs work well with independence.

11 Lecture Note(11.30)

When you have a hybrid random variable, the appropriate ways to represent the probability distribution is either with a CDF or a generalized PDF:

$$f_x(x) = f_c(x) + \sum_j a_j \delta(x - x_j)$$

Don't write something like:

$$f_X(x) = \begin{cases} kx^{-1.5} \\ 0.4 \end{cases}$$

Hw 5 is due Friday, December 7.

11.1 Look back hw2 trapping the lizards

What is the mean and standard deviation of the number of lizard trapped the second week? It seems like it would be helpful to know how many lizards are trapped the first week, and organize the calculation by conditioning on the information. (Common technique in stochastic/probability models with flow of time or sequence of events)
Let's call X_1 the number of lizards trapped the first week, and X_2 the number of lizards trapped the second week.
From the homework problem, the probability for a walking lizard to be caught in some trap during a week is

$$\tilde{p} \equiv 1 - (1-p)^t$$

$$X_1 \sim Bin(k, \tilde{p})$$

$$X_2 | X_1 \sim Bin(k - X_1, \tilde{p})$$

Or in more concrete term:

$$p^{(X_2|X_1)}_{(x_2|x_1)} = \binom{k-x_1}{x_2} \tilde{p}^{x_2}(1-\tilde{p})^{k-x_1-x_2} \text{ for } x_2 = 0, 1, 2 \cdots k - x_1$$

So

$$\begin{aligned} \mathbb{E}X_2 &= \mathbb{E}[\mathbb{E}[X_2|X_1]] \\ &= \mathbb{E}[(k - X_1)\tilde{p}] \\ &= k\tilde{p} - \tilde{p}\mathbb{E}X_1 \\ &= k\tilde{p} - \tilde{p}(k\tilde{p}) \\ &= k\tilde{p}(1-\tilde{p}) \end{aligned}$$

(btw the law of total expectation can be iterated: for number of lizards X_3 trapped in third week)

$$\mathbb{E}X_3 = \mathbb{E}[\mathbb{E}[X_3|X_2]] = \mathbb{E}[\mathbb{E}[\mathbb{E}[X_3|X_2, X_1]|X_1]]$$

11.1.1 Mean

Repeating the calculation for the mean using the more concrete approach:

$$\mathbb{E}X_2 = \sum_{x_1=0}^{k} \mathbb{E}[X_2|x_1 = x_2]p_{x_1}^{(X_1)}$$

Since the pmf of X_2 given $X_1 = x_1$ is binomial with $k - x_1$ trials and success probability p, we have:

$$\begin{aligned}
\mathbb{E}[X_2|X_1 = x_1] &= (k - x_1)\tilde{p} \\
\mathbb{E}X_2 &= \sum_{x_1=0}^{k} (k - x_1)\tilde{p}\binom{k}{x_1}\tilde{p}^{x_1}(1-\tilde{p})^{k-x_1} \\
&= ... \\
&= k\tilde{p} - k\tilde{p}^2
\end{aligned}$$

11.1.2 Standard Deviation and Variance

Standard deviation: First compute the variance:

$$\begin{aligned}
VarX_2 &= \mathbb{E}X_2^2 - (\mathbb{E}X_2)^2 \\
\mathbb{E}X_2^2 &= \mathbb{E}[\mathbb{E}[X_2^2|X_1]] \\
&= \mathbb{E}[Var(X_2|X_1) + ([X_2|X_1])^2] \\
&= \mathbb{E}[(k - X_1)\tilde{p}(1-\tilde{p}) + ((k - X_1)\tilde{p})^2] \\
&= k\tilde{p}(1-\tilde{p}) - \tilde{p}(1-\tilde{p})\mathbb{E}X_1 + \tilde{p}^2(k^2 - 2k\mathbb{E}X_1 + \tilde{p}^2\mathbb{E}X_1^2) \\
&= k\tilde{p}(1-\tilde{p}) - \tilde{p}(1-\tilde{p})(k\tilde{p}) + \tilde{p}^2(k^2 - 2k(k\tilde{p}) + \tilde{p}^2(k\tilde{p})(1-\tilde{p}) + (k\tilde{p})^2) \\
&=? \\
VarX_2 &= ...
\end{aligned}$$

But there is a shorter way. Law of total variance:

$$Var(X_2) = Var(\mathbb{E}[X_2|X_1]) + \mathbb{E}[Var(X_2|X_1)]$$

Let's redo the calculation this way:

$$\begin{aligned}
Var(X_2) &= Var((k - X_1)\tilde{p}) + \mathbb{E}[(k - X_1)\tilde{p}(1-\tilde{p})] \\
&= (-\tilde{p})^2 Var(X_1) + \tilde{p}(1-\tilde{p})(k - \mathbb{E}X_1) \\
&= \tilde{p}^2 k\tilde{p}(1-\tilde{p}) + \tilde{p}(1-\tilde{p})(k - k\tilde{p}) \\
&= \tilde{p}(1-\tilde{p})k(\tilde{p}^2 + 1 - \tilde{p})
\end{aligned}$$

Another illustration of how conditioning on an "earlier" random variable can simplify a calculation.
Let's use law of total expectation and law of total variance to provide a new derivation of the mean and standard deviation of a geometric random variable without using mgfs. The idea is when you are describing properties of a stochastic process that is memory-less or resets itself at certain moments, you can often calculate quantities through recursive formulas based on what happens up to and after the resetting moments.

Let $X \sim Geo(p)$ Note that Bernoulli trials are memory-less, so information is completely reset after each trial. Let's try conditioning on what happens until the first system reset, namely after the first trial. Let A be the event that the first trial is successful. Then obviously $\{A, A^c\}$ is a partition of state space.

11.1.3 Law of total expectation:

$$\begin{aligned}\mathbb{E}X &= \mathbb{E}[X|A]Pr(A) + \mathbb{E}[X|A^c]Pr(A^c)\\ \mathbb{E}[X|A] &= 0\\ \mathbb{E}[X|A^c] &= ?\end{aligned}$$

By the resetting property, $X|A^C \sim X + 1$

$$\begin{aligned}\mathbb{E}[X|A^c] &= \mathbb{E}[X+1]\\ &= \mathbb{E}X + 1\\ \mathbb{E}X &= 0p + (\mathbb{E}X + 1)(1-p)\\ &= (1-p)\mathbb{E}X + (1-p)\\ 0 &= -p\mathbb{E}X + (1-p)\\ \mathbb{E}X &= \frac{1-p}{p}\end{aligned}$$

11.1.4 Law of total variance:

$$Var(X) = Var(\mathbb{E}[X|I_A]) + \mathbb{E}[(Var(X|I_A))]$$

Where we have the indicator function $I_A = 1$ if A is true, else $I_A = 0$

$$\begin{aligned}\mathbb{E}[X|I_A] &= 0 && \text{if } I_A = 1\\ &= \mathbb{E}X + 1 = \frac{1}{p} && \text{if } I_A = 0\\ &= \frac{1}{p}(1 - I_A)\end{aligned}$$

$$Var(X|I_A) = ?$$

Since $X|A = 0$, so $Var(X|A) = 0$
$X|A^C \sim X + 1$ so $Var(X|A^c) = Var(X)$

$$\begin{aligned}Var(X|I_A) &= Var(X|A) && \text{if } I_A = 1\\ &= Var(X|A^c) && \text{if } I_A = 0\end{aligned}$$

$$
\begin{aligned}
Var(X|I_A) &= 0 && \text{if } I_A = 1 \\
&= Var(X) && \text{if } I_A = 0 \\
Var(X|I_A) &= Var(X)(1 - I_A) \\
Var(X) &= Var(\mathbb{E}[X|I_A]) + \mathbb{E}(Var(X|I_A)) \\
&= Var(\frac{1}{p}(1 - I_A) + \mathbb{E}[Var(X)(1 - I_A)]) \\
&= (-\frac{1}{p})^2 Var(I_A) + Var(X)(1 - \mathbb{E}I_A) \\
&= (-\frac{1}{p})^2 p(1-p) + Var(X)(1-p) \\
Var(X) &= \frac{1-p}{p^2}
\end{aligned}
$$

That calculation may seem a bit involved, but it only required manipulations of statistics of elementary random variables, and the idea generalizes to more complicated settings

11.2 Joint and Conditional Probability Distribution for Countinuous Random Variables

Essentially all concepts that we developed for multiple discrete rvs carry over to continuous rvs, just by replacing PMF concepts with PDF concepts, and integrating over the range of possible values rather than
The only technical complication, in principle, is in the conditional distribution of one continuous random variable w.r.t another continuous random variable: We would like to write down the conditional PDF of X given Y as:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \text{joint PDF/marginal PDF}$$

and that works. But its interpretation is delicate.

$$f_{X|Y}(x|y) = \lim_{\epsilon \to 0, \delta \to 0} \frac{\Pr(|X - x| < \delta | Y - y| < \epsilon}{2\delta}$$

That is, it is the probability density for X being close to the value x, given that Y is close to the value y.
Ususlly, but not always, we in fact can write this as:

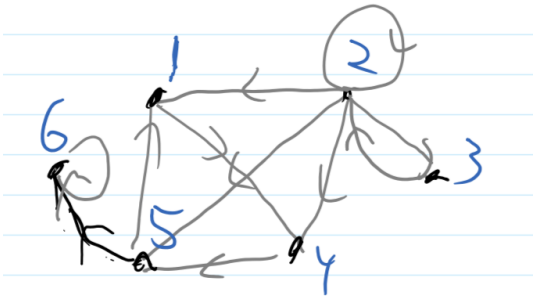
$$f_{X|Y}(x|y) = \lim_{\delta \to 0} \frac{\Pr(|X - x| < \delta | Y - y|}{2\delta}$$

That is, can we treat the event that Y is close to y by just acting as if Y=y. In practice this is almost always true–just need that the event or random variable X in which you are interested is not much affected by whether $y \approx y$ or $Y = y$. There are counterexamples in the book, see the Borel paradox in a problem in DeGroot & Schervish Sec.5.10

11.3 Markov Chains

Let's now do some example calculations involving exponentially distributed random variables. The calculations we will do have important application to Continuous-Time Markov Chains, which are systems where the state of the system can change according to various "reaction channels" and each of these reaction channels has some rate associated to it. The rate can depend on the state of the system. Continuous-Time Markov chain modeling assumption is that the system is completely memoryless, other than knowing its current state:

- chemical reaction
- atomic transitions
- polymer growth
- ecosystems
- network states(social /computer)



Two kinds of dynamic models:

- Discrete-time Markov chain: Time is a discrete sequence of updates, and at each new time, you prescribe the probability to go from one state to the other states.
- Continuous- time Markov chain: Let time flow continuously, and prescribe rates of change for moving from one state to another (these are interpreted as one over the average time the transition would take to happen.)

For continuous-time Markov chains, Memoryless assumption implies that each of the possible reactions would occur after an exponentially distributed amount of time with a mean τ_i corresponding to the rate $r_i = \frac{1}{\tau_i}$ of that particular reaction. But if there are multiple possible reactions, then when one reaction happens, it can interfere with another action. So the dynamics of the system are governed by whatever reaction happens first, then you update the state of the system, and then start over from the new state(because of memoryless property). From this perspective, the simulation of Continuous-Time Markov chain models amounts to the following(Gillespie method, kinetic Monte Carlo)

- List the possible reactions from the current state, associate a rate r_i to reaction i and/or a mean time $\tau_i = \frac{1}{r_i}$ corresponding to the average length of time you would have to wait for that particular reaction to occur.
- Generate for each of the reactions $i = 1, 2, ..., m$ an independent exponentially distributed random variable X_i with mean τ_i corresponding to the length of time you would have to wait, in this particular case for that reaction to occur
- Find the smallest value T...

- Correspondingly find the reaction whose time was first; that's the reaction that actually happens
$$J = \text{argmin}\{X_i\}_{i=1}^{m}$$
- Then, update the simulation by advancing forward by a time T, implementing reaction J to update the state of the system, then repeat the cycle

Let's calculate the probability distribution for T and J. For simplicity, we will just consider the case of m=2 reations, but the results generalize to arbitrary m, as you can read in for example Bertsekas Problem 3.39

1. Calculate the probability distribution of $T = \min(X_1, X_2)$ where X_1, X_2 are independent exponentially distributed rvs with means τ_1, τ_2 Let's use a CDF method, thinking of

$$\begin{aligned}
T &= g(X_1, X_2) \\
F_T(t) &= \Pr(T \leq t) \\
&= \Pr(\min(X_1, X_2) \leq t) \\
&= 1 - \Pr(\min(X_1, X_2) > t) \\
&= 1 - Pr(X_1 > t, X_2 > t) \\
&= 1 - \Pr(X_1 > t) \Pr(X_2 > t) \text{ Since,X1, X2 independent} \\
\Pr(X_1 > t) &= 1 - \Pr(X_1 \leq t) \\
&= 1 - F_{x_1}(t) \\
&= 1 - (1 - e^{\frac{t}{\tau_1}}) \\
&= e^{-\frac{t}{\tau_1}} \\
\Pr(X_2 > t) &= e^{-\frac{t}{\tau_2}} \\
F_T(t) &= 1 - e^{-\frac{t}{\tau_1}} e^{-\frac{t}{\tau_2}} \\
F_T(t) &= \begin{cases} 1 - e^{-\frac{t}{\bar{\tau}}} & \text{for } t \geq 0 \\ 0 & \text{for } t < 0 \end{cases} \\
\bar{\tau} &= (\tau_1^{-1} + \tau_2^{-1} text Note this is always smaller than \tau_1 \tau_2.
\end{aligned}$$

   We see from the CDF that $T \sim Exp(\bar{\tau})$

2. Now let's calculate the probability distribution of $J = argmin(X_1, X_2)$

$$\Pr(J = 1) = \Pr(X_1 < X_2) = ...?$$

   Two ways to proceed

   (a) just calculate this as an event on the sample space generated by X_1, X_2: If $f_{x_1,x_2}(x_1, x_2)$ is the joint PDF of X_1, X_2 Then

$$\Pr(X_1 < X_2) = \int_{(x_1,x_2) \in \mathbb{R}^2 : x_1 < x_2} f_{X_1,X_2}(x_1, x_2) dx_1 dx_2$$

(b) Or do a law of total probability conditioning on X_1

$$\begin{aligned}
\Pr(x_1 < X_2 | X_1 = x_1) &= \Pr(x_1 < X_2) \\
&= 1 - \Pr(X_2 \leq x_1) \\
&= 1 - F_{X_2}(x_1) \\
&= 1 - (1 - e^{-\frac{x_1}{\tau_2}}) \\
&= e^{-\frac{x_1}{\tau_2}} \\
\text{So } \Pr(X_1 < X_2) &= \tau_1^{-1} \int_0^\infty e^{-\frac{x_1}{\tau_2}} e^{-\frac{x_1}{\tau_1}} dx_1 \\
&= \tau_1^{-1} \int_0^\infty e^{-\frac{x_1}{\bar{\tau}}} dx_1 \\
&= \tau_1^{-1} \bar{\tau} \\
&= \frac{\tau_1^{-1}}{\tau_1^{-1} + \tau_2^{-1}} \\
\text{Thus, either way, } \Pr(J = 1) &= \frac{\tau_1^{-1}}{\tau_1^{-1} + \tau_2^{-1}} \\
\Pr(J = 2) &= \frac{\tau_2^{-1}}{\tau_1^{-1} + \tau_2^{-1}}
\end{aligned}$$

That is, the probability for X_i to be the smallest random variable is inversely proportional to its average. In the context of continuous-time Markov chain, the probability that the next state change will occur of along a particular reaction channel is proportional to its rate.

One can also show through a longer calculation that T,J are independent random variables. The above results generalize directly to $m > 2$ indepdendent exponentially distributed random variables. Therefore, it does seem feasible to simulate T,J directly. It is arguable, depending on application, whether simulating T,J directly is faster than simulating the m independent exponential random variables $\{X_i\}_{i=1}^m$. The point is that simulating J can be expensive due to the need to construct and sample from its probability distribution. See discussion of first reaction method and next reaction method for more on this

12 Lecture Note(12.4) Covariance and Correlation

12.1 Covariance and Correlation

Last few lectures we talked about how to do calculations involving multiple random variables. Key ideas:

- Do probability and expectation calculations by iterative conditioning so you only have to work with one random variable at a time, with the others treated as known.
- IF that does not seem to simplify the problem, think to work directly with the joint PMF/PDF/CDF
  - $$Pr((X_1, X_2, \cdots, X_n) \in A) = \sum_{(x_1,x_2,...,x_n)\in A} p)x_1, x_2, ..., x_n$$

    For discrete case; similar formulas for continuous or hybrid.
  - LUS for multiple discrete random variables:
  - $$\mathbb{E}g(X_1, X_2, \cdots, X_n) = \sum_{(x_1,x_2,...,x_n)} g(x_1, x_2, ..., x_n)p_{(x_1,x_2,...,x_n)}$$

    And similarly for continuous, hybrid rvs.

Condition probability distribution give complete description of how one random variable affect another, so does the joint probability distribution. But these are essentially 2-dimensional function(for 2 random variables) and n-dimensional functions (for n random variable), which can be cumbersome to display or sometimes even work with, just as dealing with full probability mass function is sometimes more detail than desired

For a single random variable, the mean and standard deviation were two summary statistics to condense the information in the full probability distribution.

For multiple random variables, we similarly want a simple descriptor of their relationship to each other without have to depley

12.2 Covariance of a pair of random variable X,Y:

$$\begin{aligned} Cov(X, Y) &= \mathbb{E}((X - \mathbb{E}X)(Y - \mathbb{E}Y)) \\ &= \mathbb{E}(XY) - (\mathbb{E}X)(\mathbb{E}Y) \end{aligned}$$

If X, Y are independent that $Cov(X, Y) = 0$ because then the expectation of the product can be written as the product of the expectations of each fluctuation, and these are each 0. Also $Cov(X, X) = Var(X)$

If $Cov(X, Y) > 0$, this corresponds to a "positive correlation"between the random variables X,Y meaning that an upward fluctuation in X tends to be associated with an upward fluctuation in Y. Similarly for downward fluctuations. In other words, a scatter plot of Y vs. X would have a best fit line with positive slope.

If $Cov(X,Y) < 0$, this corresponds to a "negative correlation"between the random variables X,Y meaning that an upward fluctuation in X tends to be associated with an downward fluctuation in Y. And for downward fluctuation in X tends to be associated with an upward fluctuation in Y
Note that while independent random variable have zero co-variance, one can easily construct random variables that have zero co-variance but are not independent
Two random variables with zero co-variance are side to uncorrelated.

See Interactive Scatter-plot applet for exploring how co-variance and other statistical properties of two random variables are related to data values. Related to the co-variance is the Pearson correlation coefficient:

$$\rho_{X,Y} = \rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}$$

This normalizes the co-variance to five a non-dimensional measure of the linear relation between the random variables eg $-1 \leq p_{X,Y} \leq 1$.
Pearson correlation coefficient, when you square it, it gives you thr R^2 measure of goodness of a linear fit of Y vs X.
Some useful formulas Variance of a linear combination of random variables(not necessarily independent!):

$$Var\left(b + \sum_{j=1}^{n} c_j X_j\right) = \sum_{j=1}^{n} c_j^2 Var(X_j) + 2\sum_{j=1}^{n}\sum_{j'>j} c_j c_{j'} Cov(X_j, X_{j'})$$

alternatively:

$$Var(b + \sum_{j-1}^{n} c_j X_j) = \sum_{j=1}^{n} c_j^2 Var(X_j) + \sum_{j=1}^{n}\sum_{j'\neq j} c_j c_{j'} Cov(X_j, X_{j'})$$

This extends to bi-linearity of co-variance:

$$Cov(b + \sum_{j=1}^{n} c_j X_j, \tilde{b} + \sum_{j'=1}^{m} \tilde{c_{j'}} Y_{j'}) = \sum_{j=1}^{n}\sum_{j'=1}^{m} c_j \tilde{c_{j'}} Cov(X_j, Y_{j'})$$

Let's illustrate the variance of sum formula to compute the variance of the hyper-geometric distribution.
Recall that if X is a hyper-geometric random variable, based on making a selection of k items without replacement from a population of n items, which has two subpopulations, with sizes n_1 and n_2, then we can write:
$X = \sum_{j=1}^{k} X_J$ where X_j is a Bernoulli random variable that =1 if the jth sample (with ordered sampleing) is an item from the first subpopulation, and is 0 otherwise
We already computed the mean using this setup. Now, we compute the variance. We'll use the variance of sum formula, but we need to know the variance and covariance of the X_j
Variance

$$\begin{aligned} Var(X_j) &= \Pr(X_j = 1)\Pr(X_j = 0) \\ &= \frac{n_1}{n}\frac{n-n_1}{n} \end{aligned}$$

Co-variance

$$\begin{aligned} Cov(X_j, X_{j'}) &= \mathbb{E}(X_j X_{j'}) - \mathbb{E}X_j \mathbb{E}X_{j'} \\ \mathbb{E}X_j &= \Pr(X_j = 1) = \frac{n_1}{n} = \mathbb{E}X_{j'} \\ \mathbb{E}(X_j, X_{j'}) &= ? \end{aligned}$$

At least two way to proceed. The conditional approach works if you believe in symmetry. Suppose $j < j'$.

$$\begin{aligned} \mathbb{E}(X_j X_{j'}) &= \mathbb{E}[X_j X_{j'} | X_j = 1] \Pr(X_j = 1) + \mathbb{E}[X_j X_{j'} | X_j = 0] \Pr(X_j = 0) \\ &= \mathbb{E}[1 X_{j'} | X_j = 1] \Pr(X_j = 1) + \mathbb{E}[0 X_{j'} | X_j = 0] \Pr(X_j = 0) \end{aligned}$$

First line: Law of total expectation.
Second line: Use the given information to simplify what is to the left of the bar, but don't discard the condition!

$$\begin{aligned} \mathbb{E}(X_j X_{j'}) &= \mathbb{E}[X_{j'} | X_j = 1] \Pr(X_j = 1) + 0 \\ &= \Pr(X_{j'} = 1 | X_j = 1) \Pr(X_j = 1) \end{aligned}$$

Because $\mathbb{E}X = Pr(X = 1)$ when X is Bernoulli.

$$\Pr(X_j) = 1 \frac{n_1}{n}$$

$$\Pr(X_{j'} = 1 | X_j = 1) \Pr(X_j = 1) = \frac{n_1 - 1}{n - 1}$$

So by this conditioning approach:

$$\mathbb{E}(X_j X_{j'}) = \frac{n_1 - 1}{n - 1} \times \frac{n_1}{n}$$

The slick version:

$$\mathbb{E}[X_j X_{j'}] = \mathbb{E}[\mathbb{E}[X_j X_{j'} | X_j]] = \mathbb{E}[X_j \mathbb{E}[X_{j'} | X_j]]$$

From

$$\Pr(X_{j'} = 1 | X_j = 1) \Pr(X_j = 1) = \frac{n_1 - 1}{n - 1}$$

And $Pr(X_{j'} = 1 | X_j = 0) = \frac{n_1}{n-1}$

$$\mathbb{E}[X_{j'} | X_j] = \frac{n_1}{n - 1} - \frac{1}{n - 1} X_j$$

$$
\begin{aligned}
\mathbb{E}[X_{j'}X_j] &= \mathbb{E}[X_j(\frac{n_1}{n-1} - \frac{1}{n-1}X_j)] \\
&= \frac{n_1}{n-1}\mathbb{E}X_j - \frac{1}{n-1}\mathbb{E}X_j^2 \\
\mathbb{E}X_j &= \Pr(X_j = 1) \\
&= \frac{n_1}{n} \\
\mathbb{E}X_j^2 &= 1^2 \Pr(X_j = 1) + 0^2 \Pr(X_j = 0) \\
&= \Pr(X_j = 1) \\
&= \frac{n_1}{n} \\
\mathbb{E}[X_j X_{j'}] &= \frac{n_1}{n-1}\frac{n_1}{n} - \frac{1}{n-1}\frac{n_1}{n} \\
&= \frac{n_1^2 - n_1}{n(n-1)} \\
&= \frac{n_1(n_1-1)}{n(n-1)}
\end{aligned}
$$

But in both approaches, the argument for calculating

$$Pr(X_{j'} = 1 | X_j = 1) = \frac{n_1 - 1}{n-1}$$

Might be a bit obscure.
Let's compute an alternative way, this time using joint PMF

$$
\begin{aligned}
\mathbb{E}(X_j X_{j'}) &= \sum_{x_j=0}^{1}\sum_{x_{j'}=0}^{1} x_j x_{j'} p_{x_j,x_{j'}}^{(X_j,X_{j'})} \\
&= 0 + 0 + 0 + 1 \times 1 \times p_{1,1}^{(x_j,x_{j'})} \\
p_{1,1}^{x_j,x_{j'}} &= \Pr(X_j = 1 \cap X_{j'} = 1)
\end{aligned}
$$

We can just calculate this using classical probability with a sample space corresponding to un-ordered selections of k objects without replacement from a population of size n.

$$
\begin{aligned}
\Pr(X_j = 1 \cap X_{j'} = 1) &= \frac{|X_j = 1 \cap X_{j'} = 1|}{|S|} \\
&= \frac{\binom{n_1}{2}\binom{n-2}{k-2}}{\binom{n}{k}} \\
&= \frac{\frac{n_1(n_1-1)}{2}\frac{(n-2)!}{(n-k)!(k-2)!}}{\frac{n!}{k!(n-k)!}} \\
&= \frac{n_1(n_1-1)}{2}\frac{k!(n-2)!}{(k-2)!n!} \\
&= \frac{n_1(n_1-1)k(k-1)}{n(n-1)} \quad ******
\end{aligned}
$$

Let's continue with the correct result from the iterated conditioning calculation:

$$\begin{aligned}
\mathbb{E}[X_j X_{j'}] &= \frac{n_1 - 1}{n-1} \times \frac{n_1}{n} \\
Cov(X_j, X_{j'}) &= \mathbb{E}(X_j X_{j'}) - \mathbb{E}X_j \mathbb{E}_{j'} \\
&= \frac{n_1 - 1}{n-1} \times \frac{n_1}{n} - \frac{n_1}{n}\frac{n_1}{n} \\
&= \frac{n(n_1-1)n_1 - n_1^2(n-1)}{n^2(n-1)} \\
&= \frac{nn_1^2 - nn_1 - n_1^2 n + n_1^2}{n^2(n-1)} \\
&= \frac{n_1^2 - nn_1}{n^2(n-1)} Cov(X_j, X_{j'}) \qquad\qquad = -\frac{n_1(n-n_1)}{n^2(n-1)}
\end{aligned}$$

Negative correlation, which makes sense because selecting special objects one draw destructively affects whether special object is drawn on another draw.

Coming back to the hypergeometric random variable $X = \sum_{j=1}^{k} X_j$

$$\begin{aligned}
Var(X) &= Var(\sum_{j=1}^{k} X_j) \\
&= \sum_{j=1}^{k} Var(X_j) + 2\sum_{j=1}^{k}\sum_{j'>j}^{k} Cov(X_j, X_{j'}) \qquad = \sum_{j=1}^{k} \frac{n_1}{n}\frac{n-n_1}{n} + 2\sum_{j=1}^{k}\sum_{j'>j}^{k} -\frac{n_1(n-} \\
&= k \cdot \frac{n_1}{n}\frac{n-n_1}{n} + 2 \cdot \binom{k}{2} - \frac{n_1(n-n_1)}{n^2(n-1)} \\
&= k \cdot \frac{n_1}{n}\frac{n-n_1}{n} + 2 \cdot \frac{k(k-1)}{2} - \frac{n_1(n-n_1)}{n^2(n-1)} \\
&= \frac{kn_1(n-n_1)(n-1) - k(k-1)n_1(n-n_1)}{n^2(n-1)} \\
&= \frac{n_1(n-n_1)(k(n-1) - k(k-1))}{n^2(n-1)} \\
&= \frac{n_1(n-n_1)(kn - k - k^2 + k))}{n^2(n-1)} \\
&= \frac{n_1(n-n_1)(kn - k^2)}{n^2(n-1)} \\
&= \frac{n_1(n-n_1)k(n-k)}{n^2(n-1)}
\end{aligned}$$

$Var(X)\frac{n_1(n-n_1)k(n-k)}{n^2(n-1)}$

13 Lecture Note(Dec.7)

- Bernoulli trials
  - compute probabilities of events associated to Bernoulli trials, particular numbers of success ...
- Poisson process
- Properties and modeling applications of the following standard prbability
  - Discrete uniform
  - Bernoulli/Boolean
  - binomial
- Calculations with single random variables (discrete, continuous, hybrid)
- Multiple random variables
  - sums of random variables
    - expectation of a sum is always the sum of the expectations
    - *

13.1 Bivariate normal distribution

$$f_{X_1,X_2}(x_1,x_2) = \frac{1}{2\pi\sqrt{(1-\rho^2)\sigma_1^2\sigma_2^2}} e^{\frac{-\frac{(x_1-\mu_1)^2}{\sigma_1^2}+2\frac{\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}-\frac{(x_2-\mu_2)^2}{\sigma_2^2}}{2(1-\rho^2)}}$$

ρ is correlation coefficient(see previous class)

is defined to the joint PDF for bivariate normal/Gaussian random variables (X_1, X_2), with state space $S = \mathbb{R}^2$.

The functional form looks complicated; actually the expression for the joint distribution for n random variables in terms of matrices is much easier to understand, but that's beyond the scope of the class.

The important properties bivariate normal distribution are:

- the marginal distributions are normal, in particular $X_1 \sim N(\mu_1, \sigma_1^2)$
- The only parameter appearing in the joint PDF is ρ, which is the Pearson correlation coefficient of X_1 and X_2
  - Note that when rho=0 in a bivariate distribution, then X1 and X2 is independent. Uncorrelated normal/ Gaussian random variables are independent(But this is not true for general random variables.)
- in particular, the bivariate (and in general multivariate) normal distribution is completely determined by knowing the means, variances. and co-variances(or correlations) between the set of random variables in question.
  - And in fact, one can prescribe the matrix of covariance between the random variables in any way that makes it positive semi-definite

– Multivariate distribution for n Gaussian random variables:

$$f_{X_1,\cdots X_n}(x_1\cdots x_n) = \frac{1}{\sqrt{(2\pi)^n \det C}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})\cdot C^{-1}(\vec{x}-\vec{\mu})}$$

where $\vec{\mu} = (\mathbb{X}_1, \cdots \mathbb{X}_n), \vec{x} = (x_1, x_2 \cdots x_n)$
And C is the covariance matrix, and $n \times n$ matrix whose entries are

* $C_{ij} = Cov(X_i, X_j)$

*

$$For n = 2 : c = \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

– This property is hard to extend to other probability distributions, which is why the Gaussian copula approach is often used to model multiple random variables which are not Gaussian and not independent.

Some other features of bivariate Gaussian random variables, that can be seen on the Bivariate Normal Experiment Applet:

The best-fit line through the data (X_1, X_2) generated by a bivariate normal is:

$$\mathbb{E}[X_2|X_1] = \mu_2 + \frac{\rho\sigma_2}{\sigma_1}(X_1 - \mu_1)$$

which is describing the expected value of X_2 as a function of X_1.
This can be expressed more simply in terms of "z-scores"

$$\frac{\mathbb{E}[X_2|X_1] - \mu_2}{\sigma_2} = \rho\frac{X_1 - \mu_1}{\sigma_1}$$

Also, the scatter of points about this best-fit line are given by the conditional variance formula:

$$bVar[X_2|X_1] = (1 - \rho^2)\sigma_2^2$$

If already known X1, and rho=1, the variance of X2 is 0 This expresses how the variance (as a measure of uncertainty) of X_2 is reduced by knowing the value of X_1.
These formulas are very close to what is seen in linear regression; linear regression essentially assumes that the noise in your data is joint normal.

- ρ is directly to the R^2 measure of goodness of fit

More generally, there are concepts of nonlinear regression of X_2 vs X_1, and the abstract mathematical encoding of nonlinear regression is to try and compute $\mathbb{E}[X_2|X_1]$
One could ask the question, if I have a probability model for generating the data (X_1, X_2), what deterministic function g would $X_2 = g(X_1)$ give the best fit to the data?

Answer: $g(X_1) = \mathbb{E}[X_2|X_1]$ gives the best fit in a least-squares sense.
Proof: Claim is that $\mathbb{E}[X_2|X_1]$ is the minimize of the functional $\mathbb{E}(X_2 - g(X_1))^2$
Let $g(x_1)$ be any function other than $\mathbb{E}[X_2|X_1 = x_1]$
Then $\mathbb{E}(X_2 - g(X_1))^2 = \mathbb{E}(X_2 - \mathbb{E}[X_2|X_1] - (g(X_1) - \mathbb{E}[X_2|X_1]))^2$ Let $A = X_2 - \mathbb{E}[X_2|X_1]$ and $B = (g(X_1) - \mathbb{E}[X_2|X_1])$

$$(A - B)^2 = \mathbb{E}A^2 - 2\mathbb{E}AB + \mathbb{E}B^2$$

$$\mathbb{E}(X_2 - \mathbb{E}[X_2|X_1]^2) - \mathbb{E}((X_2 - \mathbb{E}[X_2|X_1])(g(x) - \mathbb{E}[X_2|X_1])) + \mathbb{E}(g(X_1) - \mathbb{E}[X_2|X_1])^2$$

$$\begin{aligned}
\mathbb{E}((X_2 - \mathbb{E}[X_2|X_1])(g(X_1) - \mathbb{E}[X_2|X_1])) &= 0 \\
&= \mathbb{E}[\mathbb{E}[(X_2 - \mathbb{E}[X_2|X_1])(g(X_1) - \mathbb{E}[X_2|X_1])|X_1]] \\
\mathbb{E}[(X_2 - \mathbb{E}[X_2|X_1])(g(X_1) - \mathbb{E}[X_2|X_1])|X_1] & \\
g(X_1) - \mathbb{E}[X_2|X_1] \text{ :function of } X_1 & \\
&= (g(X_1) - \mathbb{E}[X_2|X_1])\mathbb{E}[(X_2 - \mathbb{E}[X_2|X_1])|X_1] \\
&= (g(X_1) - \mathbb{E}[X_2|X_1])(\mathbb{E}[X_2|X_1] - \mathbb{E}[\mathbb{E}[X_2|X_1]|X_1]) \\
&= (g(X_1) - \mathbb{E}[X_2|X_1])(\mathbb{E}[X_2|X_1] - \mathbb{E}[X_2|X_1]) \\
\mathbb{E}(X_2 - g(X_1))^2 &= \mathbb{E}(X_2 - (\mathbb{E}[X_2|X_1])^2) - 0 + \mathbb{E}(g(x_1) - \mathbb{E}[X_2|X_1]^2) \\
&> \mathbb{E}(X_2 - \mathbb{E}[X_2|X_1])^2 \\
\text{Unless } g(X_1) &= \mathbb{E}[X_2|X_1]
\end{aligned}$$

But computing $\mathbb{E}[X_2|X_1]$ isn't so easy, especially in higher dimensions.

But if (X_1, X_2) are bivariate Gaussian, then $\mathbb{E}[X_2|X_1]$ is explicitly computable we did above, and it's linear. Extending these ideas to multiple dimensions, the optimal nonlinear fit to data generated by a multivariate Gaussian distribution is always linear.
In fact, multivariate Gaussian distribution have a rigorous linear structure that makes them relatively simple to work with:

- Any finite or countable collection of jointly Gaussian random variables $X_1, X_2, \cdots$, can be viewed as vectors in a Hilbert space where the co-variance between the random variables acts like the inner product. As a byproduct, the norm on the vector space is the standard deviation of a Gaussian random variable. (orthogonal: independent)
- Conditional expectations like $\mathbb{E}[X_k|X_1, X_2 \cdots X_m]$ act geometrically in this vector space like projections on the subspace spanned by $X_1, X_2, \cdots, X_m$
- Conditional variances express the distance squared of the random variables from the their conditional expectation

These connections between multivariate Gaussians and linear algebra basically show that common data processing procedures like linear regression and principal component analysis work best when the underlying data is well modeled by a multivariate Gaussian
One more topic tail bounds. How to analize the simple random algorithm.
Sometimes, especially in analysis of algorithms, we want to have control over how bad a random fluctuation can be, and not just know what a typical random fluctuation can be.
Two basic(not very useful) inequalities are :

- Markov inequality:
$$\Pr(X > a) \leq \frac{\mathbb{E}X}{a} \text{ if } X \geq 0$$
- Chebyshev inequality:
$$\Pr(|X - \mathbb{E}X| > a) \leq \frac{VarX}{a^2}$$

A useful corollary to Chebyshev:

$$\Pr((|X - \mathbb{E}X| > t\sigma_X)) \leq \frac{1}{t^2}$$

Proof of Markov in-equality:
If $X \geq 0$, then $\mathbb{E}X = \mathbb{E}[X|X > a]Pr(X \geq a) + \mathbb{E}[X|X \leq a]\Pr(X \leq a) \geq aPr(X \geq a) + 0$

14 Lecture Note(Dec.11)

14.1 Random algorithms:

reading: Mitzenmacher and Upfal, Probability and computing, Secs. 2.5,3.3

Let's go back to Coupon Collector Problem.

The number of draws required to collect k distinct coupons from a population of m possible coupons is:

$$X = \sum_{j=1}^{k} X_j$$

where $X_j = 1 + Z_j, Z_j \sim Geo(p_j), p_j = \frac{m-(j-1)}{m}$, $\{X_j\}$ are independent
Before we computed:

$$\begin{aligned}
\mathbb{E} &= \sum_{j}^{k} = 1\frac{m}{m-(j-1)} \\
Var(X) &= Var\left(\sum_{j=1}^{k} X_j\right) \\
&= \sum_{j=1}^{k} Var\left(X_j\right) \\
\mathrm{Var}(X_j) &= \mathrm{Var}(1+Z_j) \\
&= \mathrm{Var}(Z_j) = \frac{q_j}{p_j^2} q_j \qquad\qquad = 1p_j = \frac{j-1}{m} \\
\mathrm{Var}(X) &= \sum_{j=1}^{k} \frac{\frac{j-1}{m}}{(\frac{m-(j-1)}{m})^2} \\
&= \sum_{j=1}^{k} \frac{m(j-1)}{(m-(j-1))^2}
\end{aligned}$$

The variance tells you typical fluctuations in the run time.
We can also use the mean and variance of an algorithm run time to obtain bounds on very long runs

First, we observe (see the optional reading) that for $k = m >> 1$, then:

$$\begin{aligned}
\mathbb{E}X &\sim 2mln(m) + O(1) \\
Var(X) &\leq \frac{\pi^2 m^2}{6} \\
\sigma_X &\leq \frac{\pi m}{\sqrt{6}}
\end{aligned}$$

Using just the mean of the fact that $X \geq 0$, we can apply Markov's inequality:

$$\begin{aligned}
\Pr(X \geq a) &\leq \frac{\mathbb{E}X}{a} \\
&= \frac{2m \ln m}{a} \\
\Pr(X \geq t\mathbb{E}X) &\leq \frac{2m \ln m}{t 2m \ln m} \\
&= \frac{1}{t}
\end{aligned}$$

Weak, only used information about the mean.
Using information about the variance allows us to use the somewhat better Chebyshev inequality.

$$\begin{aligned}
\Pr(|X - \mathbb{E}X| \leq a) &\leq \frac{VarX}{a^2} \\
&\leq \frac{\frac{\pi^2 m^2}{6}}{a^2} \\
&= \frac{\pi^2 m^2}{6a^2} \\
\Pr(X \geq t\mathbb{E}X) &\leq \Pr(|X - \mathbb{E}X| \geq (t-1)\mathbb{E}X) \\
&\leq \frac{\pi^2 m^2}{6((t-1)\mathbb{E})^2} \\
&= \frac{\pi^2 m^2}{6((t-1)(2m \ln m + O(1)))^2} \frac{1/m^2}{1/m^2} \\
&= \frac{\pi^2}{6((t-1)(2 \ln m + O(1/m)))^2} \\
&\sim \frac{\pi^2}{24(t-1)^2 (\ln m)^2} \text{ as } m \to \infty
\end{aligned}$$

This is a better bound than Markov inequality, if $m = 10^6$, this bounds the probability for a run time twice as large as normal by 0.0002. But this is very conservative estimate for a fluctuation that is 32 standard deviations from the mean

But this is still a weak bound. One can get a better bound by a much simpler argument."union bound"

Let $\{Y_j\}_{j=1}^m$ denote the number of times that coupon j was selected after some number n trials.

$$\begin{aligned}
\Pr(X > n) &= \Pr(\cup_{j=1}^m \{Y_j = 0\}) \\
&\leq \sum_{j=1}^m \Pr(Y_j = 0)
\end{aligned}$$

Under no assumptions: $Pr(\cup_{j=1}^{m} A_j) \leq \sum_{j=1}^{m} Pr(A_j)$

$$\begin{aligned}
\Pr(Y_j = 0) &= \left(\frac{m-1}{m}\right)^n \\
&= (1 - \frac{1}{m})^n \\
&\sim e^{-\frac{n}{m}} \text{if} m, n \to \infty \Pr(X > n) \qquad\qquad \leq \sum_{j=1}^{m} e^{-\frac{n}{m}} \\
&= me^{-\frac{n}{m}}
\end{aligned}$$

Lets try

$$\begin{aligned}
n = t\mathbb{E}X = t(2m\ln(m) + O(1) \\
&= 2tm(\ln(m) + O(1)) \\
\Pr(X > t\mathbb{E}X) &\leq me^{(} - \frac{2tm\ln m + O(1)}{m}) \\
&= me^{(} - 2t\ln m + O(\frac{1}{m})) \sim me^{-2t\ln m} = mm^{-2t} \\
\Pr(X > 2\mathbb{E}X) &\leq m^{1-2t}
\end{aligned}$$

Random quicksort: See the text for description.
Describe its run-time properties; how many comparisons X need to be made for a list of n distinct real number? Let's start with expected runtime $\mathbb{E}X$

$$X \sum_{i=1}^{n} \sum_{j \neq 1} X_{ij}$$

Where X_{ij} is an indicator variable for whether the sorted items i and j are ever compared (No pair is never compared twice since one of them must be a pivot, and once a pivot is done, it is never compared again.)
(Let the sorted output be called $(y_1.y_2, ..., y_n)$)

$$\begin{aligned}
\mathbb{E}X &= \sum_{i=1}^{n} \sum_{j \neq 1} \mathbb{E}X_{ij} \\
&= \sum_{i=1}^{n} \sum_{j > i} \mathbb{E}X_{ij} \\
\mathbb{E}X &= \Pr(X_{ij} = 1)
\end{aligned}$$

meaning that $y_i y_j$ were ever compared
To compute this, consider the list of sorted numbers from y_i to y_j": $(y_i, y_{i+1}, ..., y_{j-1}, y_j)$; remember $i < j$.
If pivots are chosen randomly, then consider which of the numbers on this list are first chosen to be a pivot. Only if y_i or y_j are chosen as first as a pivot from this list will they ever be compared. This happens with probability $\frac{2}{j-i+1} = Pr(X_{ij}) = \mathbb{E}X_{ij}$

Expected runtime:

$$\mathbb{E} = \sum_{i=1}^{n} \sum_{j>i} \frac{2}{j-i+1} \sim 2n\ln(n);\ \text{see the optional text.}$$

Now let's suppose we wanted to worry about fluctuations in runtime
Variance of the runtime:

$$Var(X) = Var\left(\sum_{i=1}^{n} \sum_{j>i} X_{ij}\right)$$

Not at all clear that the X_{ij} are independent, so that , we'd have to compute:

$$\text{Var}\left(\sum_{i=1}^{n} \sum_{j>i} \text{Var}(X_{ij})\right) = \sum_{i=1}^{n} \sum_{j>i} Var(X_{ij}) + \sum_{i=1}^{n} \sum_{j>n} \sum_{i'=1}^{n} \sum_{j'>i',(i',j')\neq(i.j)} Cov(X_{ij}, X_{i'j'})$$

Since X_{ij} is an indicator variable:

$$\begin{aligned} \text{Var}(X_{ij}) &= \Pr(X_{ij}=1)\Pr(X_{ij}=0) \\ Cov(X_{ij}, X_{i'j'}) &= \mathbb{E}X_{ij}X_{i'j'} - (\mathbb{E}X_{ij})(\mathbb{E}X_{i'j'}) \\ &= \Pr(X_{ij}=1, X_{i'j'}=1) - (\Pr(X_{ij}=1))(\Pr(X_{i'j'}=1)) \end{aligned}$$

Once you get the variance, Chebyshev...

15 Appendix A More Reading

Digital textbook on probability and statistics

16 Appendix B Relation Within Distribution