CMI/BVR Probability notes 11

joint distribution:

Just like two variables we can talk about joint distribution of several random variables. Suppose $X_1, \cdots, X_k$ are random variables on one probability space.

We say that they have joint density $f = f(x_1, \cdots, x_k)$ if

(i) f is non-negative and

(ii) for any k intervals $J_i = (a_i, b_i); 1 \leq i \leq k$

$$P(X_i \in J_i;\ \ 1 \leq i \leq k) = \int_J f(x)dx \qquad J = J_1 \times J_2 \times \cdots \times J_k.$$

Remember here $dx = dx_1 dx_2 \cdots dx_k$ and the integral is as discussed earlier.

Thus

$$P(a_i < X_i < b_i;\ \ 1 \leq i \leq k) =$$

$$\int_{x_1=a_1}^{b_1} \int_{x_2=a_2}^{b_2} \cdots \int_{x_k=a_k}^{b_k} f(x_1, x_2, \cdots, x_k)\ dx_k \cdots dx_2 dx_1$$

Of course if the above hods then we can calculate probabilities, not only for 'boxes' like J above but also for more complicated regions. For example, even in the two dimensional case we can calculate probabilities as follows. Suppose (X, Y) has density $f(x, y)$. Let us consider the region

$$D = \{(x, y) : x^2 + y^2 \leq 1\}$$

Then

$$P\{(X,Y) \in D\} = \int_D f(x,y)dxdy.$$

Remember, integrating over D means integrating as earlier the function which is f on the set D and zero off D. Thus the above equation means

$$P\{(X,Y) \in D\} = \int_0^1 \left[\int_0^{\sqrt{1-x^2}} f(x,y)dy \right] dx$$

$$= \int_0^1 \left[\int_0^{\sqrt{1-y^2}} f(x,y)dx \right] dy$$

For another example, let

$$S = \{(x,y) : x + y \leq 10\}$$

then

$$P\{(X,Y) \in S\} = \int_S f(x,y)dxdy.$$

which means

$$P\{(X,Y) \in S\} = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{10-x} f(x,y)dy \right] dx$$

$$= \int_{-\infty}^{\infty} \left[\int_{-\infty}^{10-y} f(x,y)dx \right] dy$$

That such an equality as above holds is easy to see. Though we shall not execute, here is the idea. Using geometric facts, you can express your region, (in above examples D or S) as a countable union of rectangles plus a countable union of line segments. You use countable additivity of probability. Of course, to make this argument solid, you need some cement, sand and water to be thrown in.

Independence:

Imitating the definition we had for two variables, we make the following definition.

We say random variables $X_1, X_2, \cdots, X_k$ defined on a probability space are independent if for any intervals (a_i, b_i) $(1 \leq i \leq k)$ the following holds:

$$P(a_i < X_i < b_i;\ 1 \leq i \leq k) = \prod_i P(a_i < X_i < b_i)$$

In particular, we have the following: if X_i has density f_i for $1 \leq i \leq k$ and if they are independent then $(X_1, \cdots, X_k)$ has joint density and is given by

$$f(x_1, x_2, \cdots, x_k) = f_1(x_1) f_2(x_2) \cdots f_k(x_k).$$

This is because, as in the two dimensional case, if you integrate right side over $J = \Pi J_i$ (notation as above) then definition of integration gives you $\Pi P(X_i \in J_i)$ which equals $P(X_i \in J_i; 1 \leq i \leq k)$ by independence. This shows that f satisfies the definition for being density of $(X_1, \cdots, X_k)$.

Example: $X_1, \cdots, X_{100}$ are independent unif$(0, 1)$. They have joint density

$$f(x_1, \cdots, x_{100}) = 1 \text{ if } \ 0 < x_i < 1 \ \ \forall i$$

zero otherwise. Remember; when not specified, the density is zero.

Example: $X_1, \cdots, X_{1000}$ are independent standard normal.
Then $(X_1, \cdots, X_{1000})$ has joint density given by

$$f(x) = \frac{1}{(2\pi)^{500}} e^{-x^t x/2}; \quad x \in R^{1000}.$$

Example: Suppose $X_1, \cdots, X_{50}$ are independent exponential with parameter 8 each. Then $(X_1, \cdots, X_{50})$ has joint density

$$f(x_1, \cdots, x_{50}) = 8^{50} e^{-8 \sum x_i}; \qquad x_i > 0 \ \ \forall i.$$

Example: Here is an example of non-independent random variables. Consider random variables $X_1, \cdots, X_4$ with joint density

$$f(x_1, x_2, x_3, x_4) = 24; \quad 0 < x_1 < x_2 < x_3 < x_4 < 1$$

(What is the value of f at a point of R^4 not satisfying the above condition?) You see

$$\int_0^1 [\int_{x_1}^1 [\int_{x_2}^1 [\int_{x_3}^1 24 dx_4] dx_3] dx_2] dx_1$$

$$= 24 \int_0^1 [\int_{x_1}^1 [\int_{x_2}^1 (1 - x_3) dx_3] dx_2] dx_1$$

$$= 24 \int_0^1 [\int_{x_1}^1 \frac{(1 - x_2)^2}{2} dx_2] dx_1$$

$$= 24 \int_0^1 \frac{(1 - x_1)^3}{6} dx_1$$

$$= 1$$

Clearly, here the random variables are not independent. (why?)

Sums of independent exponentials:

I have this bulb installed just now. Its life time is $X_1 \sim \exp(\lambda)$. When it burn out, without loss of time, I immediately replace by another bulb whose lifetime is similarly $\exp(\lambda)$. This is the first replacement. If X_2 is the lifetime of this new bulb, then I need to replace at time $X_1 + X_2$. It is reasonable to assume that these random variables are independent.

Thus the problem of finding the time of second replacement boils down to finding the distribution of $X_1 + X_2$ where X_1, X_2 are independent exponentials. Let us recall the density and DF of X_1, time of first replacement.

$$f_1(x) = \lambda e^{-\lambda x}; \quad x > 0 \qquad (1a)$$

$$F_1(a) = 1 - e^{-\lambda a}. \qquad (1b)$$

Denote the corresponding density and DF for $X_1 + X_2$ by f_2 and F_2. We have for $a > 0$;

$$F_2(a) = P(X_1 + X_2 \leq a) = \int_{\{x+y \leq a\}} f_1(x) f_1(y) dy dx.$$

$$= \int_{x=0}^{a} \int_{y=0}^{a-x} \lambda e^{-\lambda x} \lambda e^{-\lambda y} dx dy$$

$$\int_0^a \lambda e^{-\lambda x} \left[1 - e^{-\lambda(a-x)}\right] dx$$

$$= 1 - e^{-\lambda a} - \lambda a e^{-\lambda a}$$

Thus

$$F_2(a) = 1 - e^{-\lambda a} - \lambda a e^{-\lambda a}. \qquad (2b)$$

and

$$f_2(x) = F_2'(x) = (\lambda x)\lambda e^{-\lambda x} \qquad (2a)$$

[Of course, nobody told me, at this stage, that $X_1 + X_2$ has density. So to justify above, you need to observe that $F_2(a)$ is indeed integral of f_2 over $(-\infty, a]$. Do it. Do not forget $f_2(x) = 0$ for $x \leq 0$.]

When do we make the third replacement? at time $X_1 + X_2 + X_3$ where X_i are independent, each $\exp(\lambda)$ In other words it is the density of $Y + X_3$ where Y has density f_2 and X_3 has density f_1. Thus

$$F_3(a) = \int_{x+y\le a} f_1(x) f_2(y) dy dx$$

$$= \int_0^a \lambda e^{-\lambda x} \left[1 - e^{-\lambda(a-x)} - \lambda(a-x)e^{-\lambda(a-x)}\right] dx.$$

$$= 1 - e^{-\lambda a} - \lambda a e^{-\lambda a} - \frac{(\lambda a)^2}{2!} e^{-\lambda a}$$

Thus

$$F_3(a) = 1 - e^{-\lambda a} - \lambda a e^{-\lambda a} - \frac{(\lambda a)^2}{2!} e^{-\lambda a} \qquad (3b)$$

and

$$f_3(x) = \frac{(\lambda x)^2}{2!} \lambda e^{-\lambda x} \qquad (3a)$$

In general the n-th replacement is made at time $X_1 + \cdots + X_n$ where each X_i is $\exp(\lambda)$ and they are independent. By induction it follows that its density and DF are given by

$$f_n(x) = \frac{(\lambda x)^{n-1}}{(n-1)!} \lambda e^{-\lambda x} \qquad (na)$$

$$F_n(a) = 1 - e^{-\lambda a} - (\lambda a)e^{-\lambda a} - \cdots - \frac{(\lambda a)^{n-1}}{(n-1)!} e^{-\lambda a}$$

or

$$F_n(a) = 1 - \sum_1^n \frac{(\lambda a)^{k-1}}{(k-1)!} e^{-\lambda a} \qquad (nb)$$

You can recognize these to be gamma densities. Thus you see one way of how gamma density arises.

Poisson Process:

Thus if you have sequence of bulbs then the time T_n of the n-th replacement has density and DF given by (na) and (nb) above. We are assuming that if X_i is the life time of the i-th bulb, then

(i) for each i; $X_i \sim \exp(\lambda)$ and

(ii) for each n, the random variables $X_1, \cdots, X_n$ are independent.

An interesting question is the following: fix a time point t. At time t you ask yourself

'how many replacements did I make so far?'

If you denote by N_t the number of replacements upto time t then clearly, the event

$$(N_t = n)$$

is same as

$$(T_n \leq t) \quad \text{but} \quad (T_{n+1} \not\leq t).$$

Thus

$$P(N_t = n) = P(T_n \leq t) - P(T_{n+1} \leq t)$$

$$= F_n(t) - F_{n+1}(t) = e^{-\lambda t}\frac{(\lambda t)^n}{n!}$$

In other words N_t is a Poisson variable,

$$N_t \sim P(\lambda t).$$

This innocent question brought us to an interesting development. We started with a sequence of independent exponential random variables. Sequence is independent means for each n, the first n are independent. These are life times of bulbs. These are complicated, in the sense, they are continuous random variables. But only countable in number.

Our random variables (N_t) are each a simple random variable, discrete; it is Poisson variable. But we have many many such variables, one for each real number $t > 0$.

This process (N_t) is called Poisson process.

We studied one random variable at a time, its density, moments etc.

We learnt handling k random variables at a time; joint density etc.

A process is a collection of random variables. Usually, the index for the collection is interpreted as time parameter (though not always so). Thus if X_i is the life time of the i-th bulb, then $\{X_n : n \geq 1\}$ is a process. These are

independent random variables, according to our assumption. Another process indexed by $n = 0, 1, 2 \cdots$ is a markov chain which we discussed earlier. Here the index set is a sequence; usually called discrete parameter.

For the Poisson process $(N_t : t \geq 0)$ the index set is all non-negative real numbers, usually called continuous parameter.

We can give precise definition of Poisson process, but we shall not enter that path. First year probability course is not the correct place to discuss such matters. Only to show you the splendours of probability and give experience with sums of exponential variables I have discussed this.

expectation and variance:

We shall now convince ourselves that all concepts and theorems that we discussed for discrete variables will hold good for variables with density as well. Actually, they hold for all, but the word 'all' does not ring a bell in our minds; only 'discrete' and 'continuous (or density)' does, at this stage.

Thus through out we assume that random variables have density. If we have more than one random variable, then we assume they have joint density. As said above all this is unnecessary, but there is no need for us to be too general. Let us get the spirit.

Let $(X_1, \cdots, X_k)$ be random variables on a probability space with density f, that is, $f(x_1, \cdots, x_k)$. Suppose φ is a function of k real variables, then we define the expectation of the random variable $Z = \varphi(X_1, \cdots, X_k)$ by

$$E(\varphi(X_1, \cdots, X_k)) = \int \varphi(x) f(x) dx$$

$$= \int \varphi(x_1, \cdots, x_k) f(x_1, \cdots, x_k) dx_1 dx_2 \cdots dx_k$$

Remember, whenever we write such integrals, we assume that they exist. Otherwise, the corresponding expectation is NOT defined.

For example

$$E(\sum X_i^2) = \int (\sum x_i^2) f(x_1, \cdots, x_k) dx_1 \cdots dx_k$$

$$E(\sin X_1 e^{X_2}) = \int (\sin x_1 \; e^{x_2}) \; f(x_1, \cdots, x_k) dx_1 \cdots dx_k$$

and so on.

a subtle point: What business do I have to define expectation in the above fashion? After all; Z is a random variable in its own right; hopefully it has a density or discrete (ignore other possibilities now) and I had already defined its expectation earlier; why I am defining again, in a seemingly different way? Is it sensible to do so?

In other words, assume that Z is discrete, then should I not consider its values, $\{z_n : n \geq 0\}$ and their corresponding probabilities $\{p_n : n \geq 0\}$ and calculate $\sum z_n p_n$; as defined long ago? In case it has a density, should I not consider the density, say $h(z)$ and calculate $\int zh(z)dz$, as defined some time back?

Yes, the point is the following. present definition and the earlier definition both give the same answer. This is a theorem and we shall not dwell on that. We take it for granted. If you recall the discrete case, we had faced similar problem and solved it rigorously. You may refer to that.

Another subtle point: You may ask; why did you mention discrete random variables above because you already assumed that we now discuss random variables having density. Yes, we did assume that. Yes $(X_1, \cdots, X_k)$ has joint density. But imagine $\varphi(x_1, \cdots, x_k)$ to be the following function: if the k-tuple has all nonnegative entries then value of φ is one otherwise φ equals minus one. Think what the random variable $Z = \varphi(X_1, \cdots, X_k)$ is. In fact there are many many possibilities about which we need not worry.

Expectation is linear: In fact if X and Y have joint density $f(x, y)$ then

$$E(X + Y) = \int (x + y) f(x, y) dxdy$$

$$= \int x f(x, y) dxdy + \int y f(x, y) dxdy$$

$$= E(X) + E(Y)$$

Unfortunately, we assumed that (X, Y) has joint density. all these are true without any condition.

similarly

$$E(cX) = cE(X).$$

Fact: If X and Y are independent (assuming all expectations exist) then

$$E(XY) = E(X)E(Y)$$

Indeed if f is density of X and g that of Y, then $h(x,y) = f(x)g(y)$ is joint density of (X,Y). Hence

$$E(XY) = \int xyf(x)g(y)dydx = \int xf(x)E(Y)dx$$

$$= E(Y)E(X).$$

Variance is defined as in the discrete case:

$$var(X) = E(X^2) - [E(X)]^2$$

This is same as $E[(X-\mu)^2]$. Indeed

$$E[(X-\mu)^2] = E[X^2 - 2\mu X + \mu^2] = E(X^2) - \mu^2.$$

For two random variables X and Y we define covariance by

$$cov(X,Y) = E(XY) - E(X)E(Y)$$

This is same as

$$E[(x-\mu)(Y-\nu)]$$

where $\mu = E(X)$ and $\nu = E(Y)$. Indeed

$$E[(X-\mu)(Y-\nu)] = E(XY - \mu Y - \nu X + \mu\nu]$$

now use linearity of expectation and simplify.

Clearly $Cov(X,X) = var(X)$.

If X, Y are independent, using $E(XY) = E(X)E(Y)$ we see

$$cov(X,Y) = 0$$

Linearity of expectation leads to

$$cov(ax, bY) = ab \;\; cov(X,Y).$$

and

$$var(aX) = a^2 var(X)$$

using definition of variance, we easily show:

$$var(\sum X_i) = \sum_i var(X_i) + 2\sum_{i<j} cov(X_i, X_j)$$

In particular if the variables are independent then variance adds up:

$$var(\sum X_i) = \sum var(X_i)$$

Chebyshev

If X is a nonnegative random variable with density and if $a > 0$ then

$$P(X > a) \leq E(X)/a$$

Indeed

$$E(X) = \int_0^\infty xf(x)dx$$

remember X is non-negative and so density $f(x) = 0$ for $x \leq 0$.

$$\geq \int_a^\infty xf(x)dx \geq a \int_a^\infty f(x)dx = aP(X \geq a)$$

leading to the stated inequality.

As in the discrete case this leads to the chebyshev inequality: If X is a random variable with mean μ and variance σ^2 then

$$P(|X - \mu| > a) \leq \frac{\sigma^2}{a^2}.$$

Indeed

$$P(|X - \mu| > a) = P(|X - \mu|^2 \geq a^2) \leq \frac{\sigma^2}{a^2}.$$

WLLN

Let $X_1, X_2, \cdots$ be random variables on a probability space. Suppose they have the same distribution; they have mean μ and finite variance. Suppose that for each n, the variables $X_1, \cdots, X_n$ are independent. Then for each $\epsilon > 0$;

$$P\{|\frac{X_1 + \cdots + X_n}{n} - \mu| > \epsilon\} \to 0$$

As earlier the practical implication is that the average gets closer and closer to μ; more precisely, the chances that the average differs from μ by more than a preassigned error becomes smaller and smaller. Thus the average is a good 'estimate' for μ.

CLT:

Suppose $X_1, X_2, \cdots$ is a sequence of random variables on a probability space such that they have the same distribution; mean μ; variance $\sigma^2 > 0$. assume that for each n, the variables $X_1, \cdots, X_n$ are independent. Then for any two real numbers $a < b$,

$$P\{a < \frac{X_1 + X_2 + \cdots + X_n - n\mu}{\sqrt{n}\ \ \sigma} < b\} \to \int_a^b \frac{1}{\sqrt{2\pi}} e^{-t^2/2} dt.$$

You must appreciate that nothing about the distribution of the random variables is assumed, just that they have positive variance. You can approximate the probability on the left by the normal integral. This theorem has other uses too.

Multivariate normal:

Let $k \geq 1$ be an integer. Let Σ be a symmetric $k \times k$ positive definite symmetric matrix. Let $\mu \in R^k$. Consider the function:

$$f(x) = \frac{1}{(2\pi)^{k/2}\sqrt{|\Sigma|}} \exp\left\{-\frac{1}{2}(x-\mu)^t \Sigma^{-1}(x-\mu)\right\}; \qquad x \in R^k$$

where $|\Sigma|$ is the determinant of Σ.

We can show that this is a density function. This is called multivariate normal density. Random variables $(X_1, \cdots, X_k)$ having this density are said to be multivariate normal or jointly normal. This is the analogue of the one dimensional density

$$f(x) = \frac{1}{(2\pi)^{1/2}\sqrt{\sigma^2}} \exp\left\{\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

It appears better to close our discussion of continuous variables now and return to discrete variables.

Branching process:

There is an organism. it gives raise to certain number of offspring (could be zero), called generation 1. These in turn give raise to offspring, call generation 2. Those in generation 2 give raise to offspring, to be called generation 3 and so on. What are the chances that the generations continue for ever?

How are they growing?

To discuss a model of Bhabha, consider a high energy particle (cosmic ray?) entering earth's atmosphere. it bumps into several particles in the earth's atmosphere and by collision creates several high energy particles, call them generation 1. These in turn bump into several things in the atmosphere and by collisions create new high energy particles, call them generation 2. And so on it goes. How do you understand this process?

In nuclear chain reactions, the process starts with a neutron released. This hits several nuclei and releases neutrons, they form generation 1. These neutrons, in turn, hit several nuclei and release more neutrons, generation 2. Will the process build up (and make the bomb explode when the neutron build up excedes a critical mass) or will it die off before reaching a critical mass?

There are things called electron multipliers, which convert weak electricity into a strong one. A few electrons that exist hit a plate and thereby release many electrons, generation 1. These, in turn, hit the next plate and release more electrons, generation 2. Thus more and more electrons are released. Will the process die off or build up and produce strong current?

There is cell which is infected. This in turn infects some cells, generation 1. These newly infected cells infect some more, called generation 2. And so on it continues. How do you understand the process?

There is one commonality in all the above processes. A particle is giving raise to some new particles and these in turn give raise and so on. Such a process is called Branching Process. To get to grips with the phenomenon, we have to understand the mechanism of off-spring production. These matters are best handled by an analytical tool, called, probability generating function.

pgf:

Suppose X is a random variable taking non-negative integer values. Takes value k with probability p_k for $k = 0, 1, 2, \cdots$. The probability generating function (pgf) of X is defined as

$$\varphi(t) = p_0 + p_1 t + p_2 t^2 + \cdots$$

Note that this converges at least for $|t| \leq 1$. Our interest is mainly for

$0 \leq t \leq 1$.

Example: X is Bernoulli. Takes values 1 and 0 with probabilities p and $q = 1 - p$. Then

$$\varphi(t) = q + pt$$

Example: $X \sim B(n, p)$. Then

$$\varphi(t) = \sum_{k=0}^{n} \binom{n}{k} p^k q^{n-k} t^k = (q + pt)^n$$

Example: X is Poisson with parameter λ

$$\varphi(t) = \sum e^{-\lambda} \frac{\lambda^k}{k!} t^k = e^{-\lambda} e^{\lambda t}$$

$$= e^{-\lambda(1-t)}.$$

Similarly you can calculate pgf of geometric and other random variables. Remember, pgf is defined only for random variables taking non-negative integer values.

This is called 'probability generating' because you can discover the probabilities $\{p_k\}$ if you know the function φ. In fact $\varphi(0) = p_0$ and $\varphi'(0) = p_1$. In general

$$\varphi^{(k)}(0) = k! p_k$$

Here superscript denotes the k-th derivative. We know from the theory of power series that this function is differentiable, any number of times, in the interval $(-1, 1)$.

gf:

One can actually define generating function for any sequence of non-negative numbers $\{a_n : n \geq 0\}$ by

$$\varphi(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3 + \cdots$$

Examples:

$$a_n = \frac{1}{n!}; \qquad \varphi(t) = e^t; \quad t \in R$$

$$a_n = 1 \ \forall n; \qquad \varphi(t) = \frac{1}{1-t}; \quad t \in (-1, 1)$$

If $a_n = n^n$ then the series defining φ does not converge for any value of t other than zero. Of course, we can still keep it as a formal power series without attaching any meaning; we need not think of it as a function.

We will mostly be interested in pgf; sometimes in gf's related to pgf s.

moments from pgf:

Let now X be a random variable (non-negative integer valued) with pgf φ. Just as we recovered probabilities via derivatives at $(t = 0)$, We can recover moments via derivatives at $(t = 1)$. Of course this has to be interpreted carefully. If φ is not differentiable at $(t = 1)$ then this statement makes no sense.

We claim

$$E(X) = \lim_{t<1;t\uparrow 1} \varphi'(t)$$

This is seen as follows. For any $t \in (0, 1)$, by general theory of power series, you know

$$\varphi'(t) = p_1 + 2p_2 t + 3p_3 t^2 + \cdots$$

Thus for all $0 < t < 1$ we see

$$\varphi'(t) \leq \sum kp_k = E(X)$$

Since φ' is increasing it has limit and so

$$\lim_{t<1;t\uparrow 1} \varphi'(t) \leq E(X) \qquad (\spadesuit)$$

Also for any n we have

$$\varphi'(t) \geq \sum_0^n kp_k t^{k-1}$$

So

$$\lim_{t<1;t\uparrow 1} \varphi'(t) \geq \sum_0^n kp_k$$

This being true for every n we see

$$\lim_{t<1;t\uparrow 1} \varphi'(t) \geq \sum_0^\infty kp_k = E(X) \qquad (\clubsuit)$$

Thus $(\spadesuit)$ and $(\clubsuit)$ complete proof the claim made.

In a similar manner we can show

$$\lim_{t<1;t\uparrow 1} \varphi^{(2)}(t) = E\{X(X-1)\}$$

In general

$$\lim_{t<1;t\uparrow 1} \varphi^{(r)}(t) = E\{X(X-1)\cdots(X-r+1)\}$$

pgf of sum:

Suppose X and Y are independent random variables taking non-negative integer values. then

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t).$$

You can see this in several ways. Here is one. Observe that for any random variable V,

$$\varphi_V(t) = \sum t^k P(V=k) = E(t^V)$$

Hence

$$\varphi_{X+Y}(t) = E(t^{X+Y}) = E(t^X t^Y) = E(t^X)E(t^Y)$$

where we have used independence of X, Y. This proves the claim.

Here is another way. Let

$$P(X=k) = a_k; \qquad P(Y=k) = b_k; \qquad k = 0, 1, 2\cdots$$

Fix a $t \in [0,1]$ consider the two series

$$a_0 + a_1 t + a_2 t^2 + a_3 t^3 + \cdots$$

and

$$b_0 + b_1 t + b_2 t^2 + b_3 t^3 + \cdots$$

Their Cauchy product is nothing but

$$(a_0 b_0) + (a_0 b_1 + a_1 b_0)t + (a_0 b_2 + a_1 b_1 + a_2 b_0)t^2 + \cdots \quad (\bullet)$$

By independence of X, Y

$$P(X+Y=k) = \sum_{j=0}^{k} P(X=j, Y=k-j)$$

$$= \sum_{j=0}^{k} P(X=j)P(Y=k-j)$$

$$= a_0 b_k + a_1 b_{k-1} + \cdots + a_k b_0$$

which is the coefficient of t^k in $(\bullet)$. Thus $(\bullet)$ is $\varphi_{X+Y}(t)$. Now Cauchy's theorem on Cauchy product of series completes the proof.

generation size pgf:

Let us fix a progeny generating function

$$\varphi(t) = p_0 + p_1 t + p_2 t^2 + p_3 t^3 + \cdots$$

This is the generating function of the number of individuals produced by one particle. We assume the following:

$(\bullet)$ generation n consists of offspring of particles in generation $(n-1)$ and not others.

$(\bullet\bullet)$ Each particle produces offspring, independent of other particles and the offspring distribution is governed by φ.

The number of fellows in generation n is denoted by X_n, for $n \geq 0$ and its pgf by φ_n. Here then is the process:

We start with one particle, generation 0. Thus $X_0 = 1$. Clearly $\varphi_0(t) \equiv t$.

Its offsprings consist of generation 1. Thus X_1 takes the value j with probability p_j. So X_1 has pgf φ and thus $\varphi_1 = \varphi$.

The total number of fellows produced by those in generation 1 consists of generation 2 and their number is X_2. Thus, if $X_1 = 5$ then X_2 consists the number of offspring of these 5 fellows. More precisely, if $Y_1, \cdots, Y_5$ are the number of offspring of these five fellows, then X_2 has the same distribution as $\sum_1^5 Y_i$. This last sum has pgf $[\varphi(t)]^5$.

On the other hand if there are 10 fellows in generation 1, then X_2 would have the same distribution as $\sum_1^{10} Y_i$ where each Y_i has the pgf φ and are independent. In other words it would have pgf $[\varphi(t)]^{10}$.

In other words $P(X_2 = k | X_1 = j)$ is just the coefficient of t^k in $[\varphi(t)]^j$

How do we find the pgf of X_2?

$$P(X_2 = k) = \sum_j P(X_2 = k, X_1 = j)$$

$$= \sum_j P(X_2 = k | X_1 = j) P(X_1 = j)$$

$$= \sum_j p_j \times \text{coefficient of } t^k \text{ in } [\varphi(t)]^j$$

$$= \text{Coeficient of } t^k \text{ in } \sum_j p_j [\varphi(t)]^j$$

$$= \text{Coeficient of } t^k \text{ in } \quad \varphi(\varphi(t))$$

Thus pgf of X_2 is given by

$$\varphi_2(t) = \sum_k [\text{Coeficient of } t^k \text{ in } \ \varphi(\varphi(t))] \quad \times t^k$$

$$= \varphi(\varphi(t))$$

It is easy to guess

$$\varphi_3(t) = \varphi(\varphi(\varphi(t)))$$

Let us define 'iterated compositions' as follows:

$$\varphi_0(t) = t; \quad \text{for } n \geq 1, \quad \varphi_n(t) = \varphi(\varphi_{n-1}(t)) = \varphi_{n-1}(\varphi(t))$$

ignoring earlier convention that φ_n denotes pgf of X_n; let us temporarily agree to define φ_n as above. Let us proceed to justify that

φ_n so defined is indeed the pgf of X_n.

What we argued above is just that this claim is true for $n = 0, 1, 2$. Suppose that this is true for $n = 0, 1, 2 \cdots, m-1$. we show it is true for $n = m$. this will then complete the proof of claim. Let us observe that if the first generation has j persons, then the m-th generation consists of the $(m-1)$-th descendants of each of these j persons. Thus the total number in the m-th generation, given $X_1 = j$ has pgf $[\varphi_{m-1}(t)]^j$.

$$P(X_m = k) = \sum_j P(X_m = k, X_1 = j)$$

$$= \sum_j P(X_m = k | X_1 = j) P(X_1 = j)$$

$$= \sum_j p_j \text{ coefficient of } t^k \text{ in } \quad [\varphi_{m-1}(t)]^j$$

$$= \text{coefficient of } t^k \text{ in } \sum_j p_j [\varphi_{m-1}(t)]^j$$

$$= \text{coefficient of } t^k \text{ in } \quad \varphi([\varphi_{m-1}(t)])$$

$$= \text{coefficient of } t^k \text{ in } \quad \varphi_m(t)$$

as claimed.

Average generation size:

Let us denote the mean progeny size of a fellow is m. That is, the expected number of offspring of a fellow is m. We assume m is finite. Thus $\varphi'(1) = m$ (more precsiely limit $\varphi'(t)$ as t increases to 1 equals m).

Thus there are 'on the average' m fellows in generation 1. Each of them has 'on the average' m children and so second generation has 'on the average' m^2. Each of these has 'on the average' m children and so third generation has 'on the average' m^3 children. This is made precise below.

$$E(X_n) = \frac{d}{dt}\varphi_n(t)\Big|_{t=1}$$

$$\varphi_n'(t) = \varphi'(\varphi_{n-1}(t))\ \varphi_{n-1}'(t).$$

Hence by induction, we see $\varphi_n(1) = m^n$. Thus

$$E(X_n) = m^n$$

You can calculate the variance of X_n too. But we shall not.

The next question is whether the generations continue or extinct at some stage. That is, whether some generation size is zero. Thus we are interested in

$$\alpha = P(X_n = 0 \text{ for some } \quad n)$$

Of course, if there is none in generation 3, then there is none in future too. In other words the events $(X_n = 0)$ are increasing. Thus

$$\alpha = \lim_n P(X_n = 0)$$

Though there is much that can be discussed about branching processes, we shall close our discussion with proving the following theorem.

Note that if $p_1 = 1$ then each fellow has exactly one offspring. since we are starting our process with one person initially; we conclude that the generations continue for ever and in fact size of each generation equals exactly one. We exclude this uninteresting case. There are other uninteresting cases too, we shall see later.

Theorem: assume $p_1 \neq 1$

1. α is the smallest solution in $[0, 1]$ of the equation $\varphi(t) = t$.
2. The equation $\varphi(t) = t$ has at most two solutions in $[0, 1]$
and $t = 1$ is a solution.
3. $\alpha = 1$ if $m \leq 1$. And $\alpha < 1$ if $m > 1$.

Thus if mean offspring is at most one, then the population is extinct sooner or later. If the mean offspring is larger than one then there is a chance of survival forever.