Probability 2 - Notes1

Review of common probability distributions

1. Single trial with probability p of success. X is the indicator random variable of the event success (so $X = 1$ if the outcome is a success and $X = 0$ if the outcome is a failure). Then $X \sim$ Bernoulli(p). $P(X=1) = p$ and $P(X=0) = q$ where $q = 1-p$. $E[X] = p$, $Var(X) = pq$.

2. Sequence of n independent trials, each with probability p of success. X counts the number of successes. Then $X \sim$ Binomial(n,p). $P(X=x) = {}^nC_x p^x q^{n-x}$ for $x = 0,1,\ldots,n$. $E[X] = np$, $Var(X) = npq$.

Binomial expansion is $(a+b)^n = \sum_{x=0}^{n} {}^nC_x a^x b^{n-x}$. If we let $a = p$ and $b = q$ this shows that $\sum_{x=0}^{n} P(X=x) = (p+q)^n = 1^n = 1$.

3. Sequence of independent trials, each with probability p of success. X counts the number of trials required to obtain the first success. Then $X \sim$Geometric(p). $P(X=x) = q^{x-1}p$ for $x = 1,2,\ldots$ $E[X] = \frac{1}{p}$, $Var(X) = \frac{q}{p^2}$.

Sum of geometric series is $\sum_{x=1}^{\infty} ar^{x-1} = \frac{a}{(1-r)}$. If we let $a = p$ and $r = q$, this shows that $\sum_{x=1}^{\infty} P(X=x) = \frac{p}{1-q} = 1$.

4. Sequence of independent trials, each with probability p of success. X counts the number of trials required to obtain the k^{th} success. Then $X \sim$Negative Binomial (k,p). $P(X=x) = {}^{x-1}C_{k-1} p^k q^{x-k}$ for $x = k, k+1, \ldots$ $E[X] = \frac{k}{p}$, $Var(X) = \frac{kq}{p^2}$.

Negative binomial expansion is just

$$(1-a)^{-k} = 1 + (-k)(-a) + \frac{(-k)(-k-1)}{2!}(-a)^2 + \frac{(-k)(-k-1)(-k-2)}{3!}(-a)^3 + \ldots = \sum_{x=k}^{\infty} {}^{x-1}C_{k-1} a^{x-k}$$

Hence if we let $a = q$ then $\sum_{x=k}^{\infty} P(X=x) = p^k(1-q)^{-k} = 1$.

5. If events occur randomly and independently in time, at rate λ per unit time, and X counts the number of events in a unit time interval then $X \sim$Poisson(λ). $P(X=x) = \frac{\lambda^x e^{-\lambda}}{x!}$ for $x = 0,1,\ldots$. $E[X] = \lambda$, $Var(X) = \lambda$.

Taylor expansion of exponential is $e^a = \sum_{x=0}^{\infty} \frac{a^x}{x!}$. Hence if we let $a = \lambda$ then$\sum_{x=0}^{\infty} P(X=x) = e^{-\lambda}e^{\lambda} = 1$.

Probability Generating Function (p.g.f)

Definition For a discrete random variable X which can only take non-negative integer values we define the probability generating function associated with X to be:

$$G_X(t) = \sum_{x=0}^{\infty} P(X = x)t^x$$

This is a power series in t. Note that $G_X(t) = E[t^X]$.

We can easily find the p.g.f. for all the common probability distributions 1-5 using the expansions given earlier. Note that the hypergeometric (covered in Probability 1) has no simple form for the p.g.f.

(1) $G_X(t) = q + pt$.

(2) $G_X(t) = \sum_{x=0}^{n} {}^nC_x (pt)^x q^{n-x} = (pt+q)^n$.

(3) $G_X(t) = \sum_{x=1}^{\infty} (pt)(qt)^{x-1} = \frac{pt}{1-qt}$.

(4) $G_X(t) = (pt)^k \sum_{x=k}^{\infty} {}^{x-1}C_{k-1} (qt)^{x-k} = \frac{(pt)^k}{(1-qt)^k}$.

(5) $G_X(t) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{(\lambda t)^x}{x!} = e^{\lambda(t-1)}$.

It is easily seen that $G_X(0) = P(X = 0)$, $G_X(1) = 1$ and $G_X(t)$ is monotone increasing function of t for $t \geq 0$.

Uses of the p.g.f.

1. Knowing the p.g.f. determines the probability mass function.

The p.g.f., $G_X(t)$, is a power series with the coefficient of t^x just the probability $P(X = x)$. There is a unique power series expansion. Hence if X and Y are two random variables with $G_X(t) = G_Y(t)$, then $P(X = r) = P(Y = r)$ for all $r = 0, 1, \ldots$.

If we know the p.g.f. then we can expand it in a power series and find the individual terms of the probability mass function.

e.g. $G_X(t) = \frac{1}{2}(1+t^2) = \frac{1}{2} + 0 \times t + \frac{1}{2}t^2 + 0 \times t^3 + \ldots$. Hence $P(X = 0) = \frac{1}{2}$, $P(X = 2) = \frac{1}{2}$ and $P(X = x) = 0$ for all other non-negative integers x.

If we recognise the p.g.f. $G_X(t)$ as a p.g.f. corresponding to a specific distribution, then X has that distribution. We do not need to bother doing the power series expansion!

e.g. if $G_X(t) = e^{2t-2} = e^{2(t-1)}$, this is the p.g.f. for a Poisson distribution with parameter 2. Hence $X \sim$Poisson (2).

2. We can differentiate the p.g.f. to obtain $P(X=r)$ and the factorial moments (and hence the mean and variance of X).

$$P(X=0)=G_X(0);\ P(X=1)=G_X^{'}(0);\ P(X=2)=\frac{1}{2}G_X^{''}(0)$$

In general $P(X=r)=\frac{1}{r!}G_X^{(r)}(0)$ where $G_X^{(r)}(t)=\frac{d^rG_X(t)}{dt^r}$.

$$E[X]=G_X^{'}(1);\ E[X(X-1)]=G_X^{(2)}(1);\ Var(X)=E[X(X-1)]+E[X]-(E[X])^2$$

and in general the r^{th} factorial moment $E[X(X-1)...(X-r+1)]=G_X^{(r)}(1)$

This is easily seen by differentiating $G_X(t)=P(X=0)+tP(X=1)+t^2P(X=2)+...$ termwise to obtain

$$G_X^{'}(t)=P(X=1)+2tP(X=2)+3t^2P(X=3)+....$$

from which we have $E[X]=G_X^{'}(1)$ and $P(X=1)=G_X^{'}(0)$ and for any positive integer r

$$\frac{d^rG_X(t)}{dt^r}=r!P(X=r)+\frac{(r+1)!}{1!}tP(X=r+1)+\frac{(r+2)!}{2!}t^2P(X=r+2)+...$$

from which we have $E[X(X-1)...(X-r+1)]=G_X^{(r)}(1)$ and $P(X=r)=\frac{G_X^{(r)}(0)}{r!}$

e.g. If $G_X(t)=\frac{1+t}{2}e^{(t-1)}$ find $E[X]$, $Var(X)$, $P(X=0)$ and $P(X=1)$.

$$G_X^{'}(t)=\frac{1}{2}e^{(t-1)}+\frac{1+t}{2}e^{(t-1)}$$

$$G_X^{(2)}(t)=\frac{1}{2}e^{(t-1)}+\frac{1}{2}e^{(t-1)}+\frac{1+t}{2}e^{(t-1)}$$

Hence $E[X]=G_X^{'}(1)=\frac{3}{2}$, $var(X)=G_X^{(2)}(1)+\frac{3}{2}-\frac{9}{4}=\frac{5}{4}$, $P(X=0)=G_X(0)=\frac{e^{-1}}{2}$ and $P(X=1)=G_X^{'}(0)=e^{-1}$.

3. Using the p.g.f. to find the distribution of the sum of two or more independent random variables.

Recall that if X and Y are independent random variables then $E[g(X)h(Y)]=E[g(X)]E[h(Y)]$.

Let X and Y be independent random variables with p.g.f.'s $G_X(t)$ and $G_Y(t)$. Then $Z=X+Y$ has p.g.f.

$$G_Z(t) = E[t^Z] = E[t^{X+Y}] = E[t^X t^Y] = E[t^X]E[t^Y] = G_X(t)G_Y(t)$$

This extends to the sum of a fixed number n of independent random variables.

If $X_1, ..., X_n$ are independent and $Z = \sum_{j=1}^{n} X_j$ then

$$G_Z(t) = \prod_{j=1}^{n} G_{X_j}(t)$$

e.g. Let X and Y be independent random variables with $X \sim Binomial(n, p)$ and $Y \sim Binomial(m, p)$ and let $Z = X + Y$. Then

$$G_Z(t) = G_X(t)G_Y(t) = (pt + q)^n (pt + q)^m = (pt + q)^{m+n}$$

This is the p.g.f. of a binomial random variable. Hence $Z \sim Binomial(n + m, p)$.

Let $X_1, \ldots, X_m$ be m independent random variables with $X_j \sim Binomial(n_j, p)$ and let $Z = \sum_{j=1}^{m} X_j$ and $N = \sum_{j=1}^{m} n_j$. Then

$$G_Z(t) = \prod_{j=1}^{m} G_{X_j}(t) = \prod_{j=1}^{m} (pt + q)^{n_j} = (pt + q)^N$$

This is the p.g.f. of a binomial random variable. Hence $Z \sim Binomial(N, p)$.