37161 Probability and Random Variables

Lecture 8

Sums of Random Variables

- Consider now two independent random variables which arise from flipping two fair coins a fixed number of times.

- If Coin A is flipped 5 times, the number of times it lands Tails is $N_A \sim Bin(5, 0.5)$.

- If Coin B is flipped 10 times, the number of times it lands Tails is $N_B \sim Bin(10, 0.5)$.

- What is the distribution of the total number of times either of the two coins land Tails, $N_A + N_B$?

Sums of Random Variables

- Intuitively, we can conclude that $\left[N_A + N_B\right] \sim Bin(15, 0.5)$ but proving this is less trivial.

- We can perhaps justify this by appealing to the fact that a binomial variable can be considered as the sum of a number of independent identically distributed Bernoulli variables.

- We could evaluate the distribution of sums of random variables via convolutions, for example

$$P\left(\left[N_A + N_B\right] = k\right) = \sum_{j=0}^{k} P(N_A = j)P(N_B = k - j)$$

$$= \sum_{j=0}^{k} \frac{5!}{(5-j)!\,j!} 0.5^{j} 0.5^{5-j} \frac{10!}{(10-(k-j))!)(k-j)!} 0.5^{k-j} 0.5^{10-(k-j)}$$

Sums of Random Variables

- $P\left([N_A + N_B] = k\right) = \sum_{j=0}^{k} \frac{5!}{(5-j)!\,j!} 0.5^{j} 0.5^{5-j} \frac{10!}{(10-(k-j))!)(k-j)!} 0.5^{k-j} 0.5^{10-(k-j)}$

- This is, however, not easy to work with and justifying that the above statement simplifies to $\frac{15!}{k!(15-k)!} 0.5^{15}$ requires knowing several identities regarding summing binomial coefficients.

- Adding more than two random variables is even messier. For example, for independent variables X_1, X_2, X_3, X_4 , the probability that $[X_1 + X_2 + X_3 + X_4] = k$ requires summing over possible ways this could happen.

$$[X_1 + X_2 + X_3 + X_4] = k = \sum_{n=0}^{k} \sum_{m=0}^{k-n} \sum_{j=0}^{k-m-n} P(X_1 = j) P(X_2 = m) P(X_3 = n) P(X_4 = k - j - m - n)$$

Generating Functions

- Quite clearly, the convolution approach is not practical for any large sum of random variables.

- Instead, many such calculations can be more easily done with **generating functions**.

- These are transformations of the probability mass function or probability density function of the variable, such that some key properties of the variable can still be recovered.

Generating Functions

- The generating function of a random variable X is defined as $g_X(z) = E(z^X)$.
- If X is discrete, $g_X(z) = E(z^X) = \sum P(X = k)z^k$.
- If X is continuous $g_X(z) = E(z^X) = \int z^x f(x)dx$.

Generating Functions: Example

- Consider rolling one regular fair six sided die.

- The probability mass function for this is $P(X=k)=\begin{cases} 1/6 & k=1 \\ 1/6 & k=2 \\ 1/6 & k=3 \\ 1/6 & k=4 \\ 1/6 & k=5 \\ 1/6 & k=6 \\ 0 & \text{otherwise} \end{cases}$

- The generating function is therefore $g_X(z)=E(z^X)=\sum P(X=k)z^k$

$$=\frac{1}{6}z^1+\frac{1}{6}z^2+\frac{1}{6}z^3+\frac{1}{6}z^4+\frac{1}{6}z^5+\frac{1}{6}z^6$$

Expectation and Variance

- If given the generating function of a variable, how do we obtain the expectation or variance of the underlying variable?

- We know that $g_X(z) = E(z^X) = \sum P(X = k)z^k$. Differentiating gives $\dfrac{dg_X(z)}{dz} = \sum P(X = k)kz^{k-1}$.

- $E(X) = \sum P(X = k)k$, so this is obtained by finding $g'_X(1) = \sum P(X = k)k(1^{k-1})$.

- This gives $E(X) = g'_X(1)$. In other words, the expectation of X can be recovered by differentiating its generating function $g_X(z)$ and evaluating when $z = 1$.

Expectation and Variance

- Similarly, differentiating a second time gives $\dfrac{d^2 g_X}{dz^2} = \sum_{k=0}^{\infty} k(k-1)P(X=k)z^{k-2}$

$$= \sum_{k=0}^{\infty} k^2 P(X=k)z^{k-2} - \sum_{k=0}^{\infty} kP(X=k)z^{k-2}$$

- Again, setting $z=1$ gives $g''_X(1) = \sum P(X=k)k^2(1^{k-2}) - \sum P(X=k)k(1^{k-2}) = E(X^2) - E(X)$

- $\mathrm{Var}(X) = E(X^2) - E(X)^2$ hence $\mathrm{Var}(X) = g''_X(1) + g'_X(1) - g'_X(1)^2$.

Generating Functions: Sums of Variables

- Returning now to the problem of summing random variables, let X_1, X_2 be independent random variables both taking non-negative integer values.

- Let. $Y = X_1 + X_2$, then $P(Y = k) = \sum_{j=0}^{k} P(X_1 = j) P(X_2 = k - j)$.

- Now, consider the generating function of Y, $g_Y(z) = \sum_{k=0}^{\infty} P(Y = k) z^k$

$$= \sum_{k=0}^{\infty} \left[\sum_{j=0}^{k} P(X_1 = j)\, P(X_2 = k - j) \right] z^k$$

Generating Functions: Sums of Variables

- $g_Y(z) = \sum_{k=0}^{\infty}\left[\sum_{j=0}^{k} P(X_1 = j)\, P(X_2 = k-j)\right] z^k$

- Since X_1 and X_2 are independent, we can split the summation

$$g_Y(z) = \sum_{k=0}^{\infty}\left[\sum_{j=0}^{k} P(X_1 = j)\, P(X_2 = k-j)\right] z^k = \sum_{k=0}^{\infty}\left[\sum_{j=0}^{k} P(X_1 = j) z^j P(X_2 = k-j) z^{k-j}\right]$$

$$= \sum_{k=0}^{\infty}\left[g_{X_1}(z) P(X_2 = k-j) z^{k-j}\right] = g_{X_1}(z) g_{X_2}(z)$$

- That is, for $Y = X_1 + X_2$, $g_Y(z) = g_{X_1}(z) g_{X_2}(z)$.

Generating Functions: Sums of Variables

- The fact that we can calculate sums of independent random variable through multiplication of their generating functions is perhaps unsurprising, since $g_Y(z) = E(z^{X_1+X_2}) = E(z^{X_1})E(z^{X_2})$..
- This also gives us simple and quick methods of verifying relationships we have already seen.
- Let $X_1,\ldots,X_n \sim Poi(\lambda)$ be n independent Poisson variables.
- Each of these has generating function $g_{X_i} = E(z^{X_i}) = \sum_{k=0}^{\infty} z^k \frac{e^{-\lambda}\lambda^k}{k!} = \sum_{k=0}^{\infty} \frac{e^{-\lambda}(z\lambda)^k}{k!}$.
- We know $\sum_{k=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = 1$ hence $\sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}$ so $\sum_{k=0}^{\infty} \frac{e^{-\lambda}(z\lambda)^k}{k!} = e^{-\lambda}(e^{z\lambda}) = e^{\lambda(z-1)}$

Generating Functions: Sums of Variables

- We now know that if $X_i \sim Poi(\lambda)$, its generating function is $g_{X_i}(z) = e^{\lambda(z-1)}$.
- So, for $Y = X_1 + \ldots + X_n$, we have $g_Y(z) = E(z^Y) = E(z^{X_1+\ldots+X_n}) = E(z^{X_1})E(z^{X_2})\ldots E(z^{X_n})$.
- Each of $E(z^{X_1}),\ E(z^{X_2}),\ldots,E(z^{X_n})$ is the same function, equal to $g_{X_i}(z) = e^{\lambda(z-1)}$.
- The generating function of Y is therefore $g_Y(z) = \left[g_{X_i}(z)\right]^n = \left[e^{\lambda(z-1)}\right]^n = e^{n\lambda(z-1)}$
- We therefore have that, if $X_i \sim Poi(\lambda)$ and each variable is independent, then $Y = X_1 + \ldots + X_n \sim Poi(n\lambda)$
- This is equivalent to (but simpler in its derivation) what we previously saw regarding merging independent Poisson processes.

Generating Functions: Bernoulli Variables

- Let T be a Bernoulli variable $T \sim Bern(p)$.

- It is easy to see directly from the probability mass function $P(T=k)=\begin{cases} 1-p & k=0 \\ p & k=1 \\ 0 & \text{otherwise} \end{cases}$

  that the generating function is $g_T(z)=E(z^T)=(1-p)z^0+pz=1-p+pz$

Generating Functions: Binomial Variables

- Let S be a binomial variable $S \sim Bin(n, p)$

- Consider $S = T_1 + ... + T_n$ where $T_1, ..., T_n \sim Bern(p)$ are independent Bernoulli variables.

- As S is the sum of n independent Bernoulli variables, its generating function is the product of the n generating functions of these variables, $g_S(z) = \left[g_{T_i}(z)\right]^n$.

- Since each $g_{T_i}(z) = 1 - p + pz$ then $g_S(z) = \left[g_{T_i}(z)\right]^n = [1 - p + pz]^n$

Generating Functions: Geometric Variables

- Let $X \sim Geo(p)$ be a Geometric random variable.
- The probability mass function is therefore $P(X = k) = \begin{cases} (1-p)^{k-1}p & k = 1,2,3,\ldots \\ 0 & \text{otherwise} \end{cases}$
- This gives the generating function $g_X(z) = E(z^X) = \sum_{k=1}^{\infty} z^k P(X = k) = \sum_{k=1}^{\infty} z^k (1-p)^{k-1} p$

$$= zp + z^2 p(1-p) + z^3 p(1-p)^2 + z^4 p(1-p)^3 + \ldots$$

- This is a geometric series with first term zp and common ratio $(1-p)z$.
- (Provided z is chosen such that $|(1-p)z| < 1$) this sums to give $g_X(z) = \dfrac{zp}{1-(1-p)z}$.

Generating Functions: Exponential Variables

- Let $Y \sim \exp(\lambda)$ be an exponential random variable.
- The probability density function is therefore $f(y) = \begin{cases} \lambda e^{-\lambda y} & y \in [0,\infty) \\ 0 & \text{otherwise} \end{cases}$
- This gives the generating function $g_Y(z) = E(z^Y) = \int_0^\infty z^y f(y)dy = \int_0^\infty z^y \lambda e^{-\lambda y} dy$
- In order to evaluate this, we need to combine $z^y e^{-\lambda y}$ into a single exponential.
- Taking the exponential of the logarithm of z^y we get $z^y = e^{\ln(z^y)} = e^{y\ln(z)}$ and hence $e^{-y(\lambda - \ln(z))}$.

Generating Functions: Exponential Variables

- $g_Y(z) = E(z^Y) = \int_0^{\infty} z^y f(y)dy = \int_0^{\infty} e^{y\ln(z)} \lambda e^{-\lambda y} dy$

- $g_Y(z) = E(z^Y) = \lambda \int_0^{\infty} e^{-y(\lambda - \ln(z))} dy \quad -\lambda \left[\dfrac{e^{-y(\lambda - \ln(z))}}{\lambda - \ln(z)} \right]_0^{\infty} \quad = \dfrac{\lambda}{\lambda - \ln(z)}$ (provided $\lambda > \ln(z)$)

- The generating function of Y is therefore $g_Y(z) = \dfrac{\lambda}{\lambda - \ln(z)}$.

Sum of a Random Number of Random Variables

- For many applications, we are interested in the sum of a number of random variables where the number of variables to be summed is itself random.

- For example, the total annual payouts for an insurance company varies according to two variables – the average size of a claim and the total number of individual claims.

- The total payout is the sum of each individual claim, summed over an uncertain number of claims.

Sum of a Random Number of Random Variables

- Consider now the problem where N is a random variable taking non-negative integer values and $X_1, X_2, ..., X_N$ are independent, identically distributed random variables.
- Let S_N be the sum of these variables i.e. $S_N = \sum_{k=1}^{N} X_k$.
- The generating function of S_N is therefore $g_{S_N}(z) = E(z^{S_N})$.
- If we knew the value of N, say $N = k$, this would be easy to evaluate. Since $X_1, X_2, ..., X_N$ all have the same distribution, we would simply be multiplying k identical generating functions $g_{S_N}(z) = \left[g_{X_i}(z)\right]^k$.

Sum of a Random Number of Random Variables

- We know by conditional expectation that\ $g_{S_N}(z) = E(z^{S_N}) = E(E(z^{S_N} \mid N = k))$

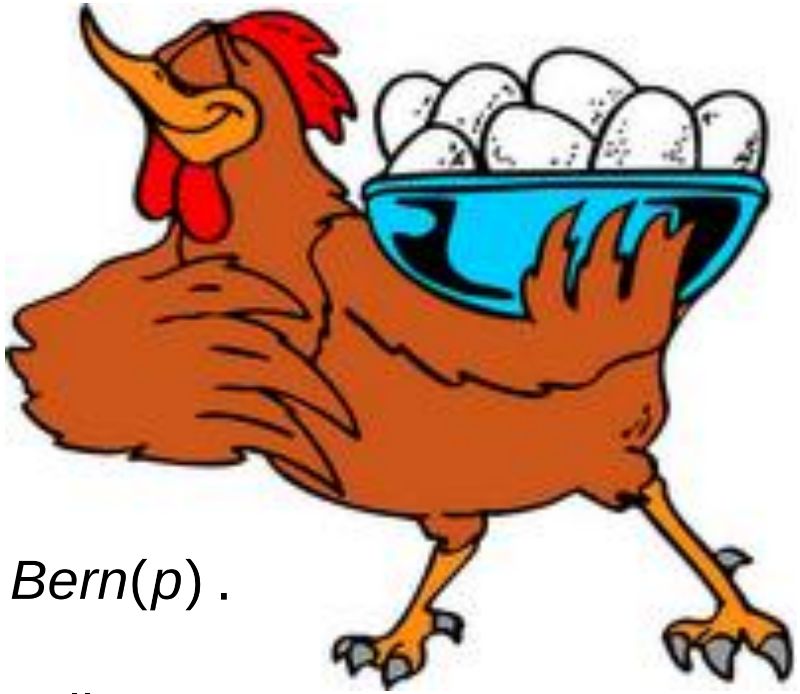
$$= \sum_{k=0}^{\infty} E(z^{S_N} \mid N = k) P(N = k)$$

- Here, this gives $g_{S_N}(z) = E(z^{S_N}) = \sum_{k=0}^{\infty} \left[g_{x_i}(z) \right]^k P(N = k)$

- Now, the last term is itself a generating function applied to a generating function. In general, for any non-negative discrete variable Q, $g_Q(z) = E(z^Q) = \sum_{k=0}^{\infty} z^k P(Q = k)$ so we therefore have that $g_{S_N}(z) = g_N(g_{x_i}(z))$.

- In other words, if we are adding N independent identically distributed variables $X_1, X_2, ..., X_N$ then the generating function of the sum is equal to the generating function of N evaluated when z equals the generating function of each of the X_i variables.

Example: Poisson Hen



- Consider a hen which lays N eggs, where $N \sim Poi(\lambda)$ and each egg hatches to produce one chicken with probability p, independent of all other eggs.
- The number of chickens from each single egg is therefore $X_i \sim Bern(p)$.
- What is the distribution of the number of chickens hatching from all eggs, S_N ?
- We already know that $g_{X_i}(z) = (1-p) + pz$ and $g_N(z) = e^{\lambda(z-1)}$
- These therefore give $g_{S_N}(z) = g_N(g_{x_i}(z)) = e^{\lambda\left[g_{X_i}-1\right]} = e^{\lambda[(1-p)+pz-1]} = e^{p\lambda(z-1)}$
- Since the generating function of S_N is $e^{p\lambda(z-1)}$, we can recognise this as a Poisson variable, $S_N \sim Poi(p\lambda)$.

Example

- Consider a car which drives past other vehicles at the instants of a Poisson process with rate parameter equal to 10 cars per minute.

- This gives that the number of cars passed in a minute $\sim Poi(10)$ and also that the time (in minutes) between successive cars passed $\sim \exp(10)$.

- What is the distribution of the time until the car next passes a blue car, assuming that each car passed is a blue car with probability 20%, independent of the colours of all other cars?

- For example, if the car passes a green car after 0.121 minutes, then a black car 0.207 minutes after that then a blue car after a further 0.088 minutes, the time until the next blue car would be recorded as $0.121+0.207+0.088=0.416$ minutes.

Example

- The total time until an additional blue car has been passed is equal to the sum of a random number of random numbers.

- We know that the time between passing successive cars $\sim \exp(10)$ but we also need to know the distribution of how many cars will be passed until passing the next blue car.

- If each car is blue with probability 20%, then we are counting how many independent identical $Bern(0.2)$ variables until the first 1 (or blue) is observed.

- We therefore have that the number of cars passed until passing a blue car $\sim Geo(0.2)$.

Example

- Let W_i be the additional waiting time until the *i*th car is passed (e.g. W_5 Is the time after the fourth car has been passed until the fifth car is passed.)

- In minutes, each $W_i \sim \exp(10)$.

- The time until the next blue car is passed is therefore

$W_{blue} = \sum_{i=1}^{X} W_i$ where X is the number of cars passed until passing a blue car.

- Here, $X \sim Geo(0.2)$.

Example

- We sum this random number of random numbers via generating functions.

- $W_i \sim \exp(10)$ hence $g_{W_i}(z) = \dfrac{10}{10 - \ln(z)}$.
- $X \sim Geo(0.2)$ hence $g_X(z) = \dfrac{0.2z}{1 - 0.8z}$.
- The generating function of $W_{blue} = \sum_{i=1}^{X} W_i$ is given by $g_{W_{blue}}(z) = g_X(g_{W_i}(z))$

Example

- $g_{W_{blue}}(z) = g_X(g_{W_i}(z)) = \dfrac{0.2\left(\dfrac{10}{10-\ln(z)}\right)}{1-0.8\left(\dfrac{10}{10-\ln(z)}\right)}$

$$= \frac{2}{10-\ln(z)-0.8(10)} = \frac{2}{2-\ln(z)}.$$

- As $g_{W_{blue}}(z) = \dfrac{2}{2-\ln(z)}$ we can see that $W_{blue} \sim \exp(2)$.

- This is equivalent to (but simpler in its derivation) what we previously saw regarding splitting Poisson processes.