**Tutorial 12.**

# 1 Sums of random variables

Sums of random variables arise in many applications. The basic question is: given a pair (or a set) of random variables, what is the distribution of the random variable defined as their sum?

For two random variables, the distribution function for their sum can be obtained by a special formula. Let the random variables be $R_1$ and $R_2$ with distribution fuctions $f_1$ and $f_2$. Then the distribution function for their sum is

$$g(x) = \sum_{x_1,x_2:x_1+x_2=x} f(x_1, x_2)$$

Here $f$ is the *joint* distribution function of $f_1$ and $f_2$, i.e., $f(x)$ is the probability that $R_1 = x_1$ and $R_2 = x_2$.

$$g(x) = \sum_{x_1,x_2:x_1+x_2=x} f_1(x_1)f_2(x_2 \mid x_1)$$

If $R_1$ and $R_2$ are independent, then

$$g(x) = \sum_{x_1,x_2:x_1+x_2=x} f_1(x_1)f_2(x_2)$$

Do you recognize this? $g$ is the *convolution* of $f_1$ and $f_2$.

We can represent this nicely in terms of generating functions in the special case when the range of $R_1$ and $R_2$ is $\mathcal{N}$. For this we define generating functions of probabilities.

**Definition 1** *Let $R$ be a random variable $R : S \rightarrow \mathcal{N}$, then the* probability generating function (PGF) *for $R$ is*

$$G(z) = \sum_{i=0}^{\infty} Pr(R = i)z^i$$

*Examples:*

- $R$ is Bernoulli: $Pr(R = 1) = p$, $Pr(R = 0) = 1 - p$.

$$G(z) = (1 - p) + pz$$

- $R$ is uniform on $[0, N - 1]$.

$$G(z) = \frac{1}{N} + \frac{1}{N}z + \frac{1}{N}z^2 + \ldots + \frac{1}{N}z^{N-1} = \frac{1}{N}\left(\frac{1 - z^N}{1 - z}\right)$$

- $R$ is binomial with parameters $p$ and $n$.

$$G(z) = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} z^k = [(1-p) + pz]^n$$

Now we can state the product rule for sums of independent random variables. Notice that this is analogous to the product rule when we discussed counting.

**Theorem 1** *If $R_1$ and $R_2$ are independent random variables from $S$ to $\mathcal{N}$ with PGF's $G_1(z)$ and $G_2(z)$, then the PGF of $R = R_1 + R_2$ is*

$$G(z) = G_1(z)G_2(z)$$

*Proof.*

$$\begin{aligned}
G(z) &= \sum_{n=0}^{\infty} Pr(R = n) z^n \\
Pr(R = n) &= Pr(R_1 + R_2 = n) \\
&= \sum_{i,j:i+j=n} Pr(R_1 = i \wedge R_2 = j) \\
&= \sum_{i,j:i+j=n} Pr(R_1 = i) Pr(R_2 = j)
\end{aligned}$$

The last step uses the fact that $R_1$ and $R_2$ are independent. But then, the coefficient of $z^n$ in $G(z)$ is

$$Pr(R = n) \quad = \quad \text{coefficient of } z^n \text{ in } \left( \sum_{i=0}^{\infty} Pr(R_1 = i) z^i \right) \left( \sum_{j=0}^{\infty} Pr(R_2 = j) z^j \right)$$

Hence $G(z) = G_1(z)G_2(z)$. ■

As an example here is a problem we've seen before, but we'll view it differently. There are $n$ devices in a circuit we are building. Each of the $n$ devices could be faulty. The probability that any one fails is $p$, and the failures are independent. ¿From this we would like to get the distribution for the number of failures (e.g., what is the probability that there are 2 failures?).

Define the random variables $R_1, \ldots, R_n$ as

$$R_i = \begin{cases} 1 & i^{th} \text{ device fails} \\ 0 & \text{otherwise} \end{cases}$$

Then,

$$\forall i, Pr(R_i = 1) = p$$

Let $R$ be the number of failures overall.

$$R = R_1 + R_2 \ldots + R_n$$

The PGF for each $R_i$ is $(1-p)+pz$. Therefore, by the product rule, the PGF for $R$ is $[(1-p)+pz]^n$. Hence, the probability of $k$ errors is

$$\begin{aligned} Pr(R=k) &= \text{coefficient of } z^k \\ &= \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

This is also an alternative derivation of the Binomial distribution.

The expectation of a random variable can be expressed compactly via its PGF.

**Theorem 2** *Let $R : S \rightarrow \mathcal{N}$ be a random variable with PGF $G(z)$.*

$$E(R) = G'(1)$$

*Proof.*

$$\begin{aligned} G(z) &= \sum_{i=0}^{\infty} Pr(R=i) z^i \\ G'(z) &= \sum_{i=0}^{\infty} Pr(R=i) i z^{i-1} \\ G'(1) &= \sum_{i=0}^{\infty} Pr(R=i) i \\ &= E(R) \end{aligned}$$

For example, the expectation of a binomial distribution with parameters $p$ and $n$:

$$G'(z) = n[(1-p)+pz]^{n-1} p$$

$$G'(1) = np$$

# 2 The Poisson distribution

Next we study the Poisson distribution which occurs in a surprising variety of real-life applications.

One setting for a Poisson distribution is random events occurring in time, such as radioactive disintegrations, incoming telephone calls or people coming to a cash register. Each event can be a point on the time axis and we are interested in the distribution of points, namely, what is the probability of finding exactly $k$ points within a fixed interval of specified length? So for example this could be the number of $\alpha$-particles emitted by a radioactive substance that reach a given portion of space during time $t$. This particular example is actually known to correspond very closely to the Poisson distribution.

The *Poisson* distribution is

$$Pr(R = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

for the random variable $R : S \to \mathcal{N}$. Here $\lambda > 0$ is a parameter of the distribution.

Let us quickly check that this indeed a valid distribution.

$$\begin{aligned}
\sum_{k=0}^{\infty} Pr(R = k) &= \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} \\
&= e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} \\
&= e^{-\lambda} e^{\lambda} \\
&= 1
\end{aligned}$$

In the final exam a few years ago, students were asked to calculate the expectation of $R$. For this let us write the PGF for $R$.

$$\begin{aligned}
G(z) &= \sum_{k=0}^{\infty} Pr(R = k) z^k \\
&= \sum_{k=0}^{\infty} \frac{\lambda^k e^{-\lambda}}{k!} z^k \\
&= e^{-\lambda} \sum_{i=0}^{\infty} \frac{(\lambda z)^k}{k!} \\
&= e^{-\lambda} e^{\lambda z} \\
&= e^{-\lambda + \lambda z}
\end{aligned}$$

Then $G'(z) = \lambda e^{-\lambda + \lambda z}$ and $E(R) = G'(1) = \lambda$.

They were also asked to show that the sum of independent Poisson distributed random variables is itself Poisson distributed.

Let $R_1, R_2$ be random variables with distributions

$$Pr(R_1 = k) = \frac{\lambda_1^k e^{-\lambda_1}}{k!}$$

$$Pr(R_2 = k) = \frac{\lambda_2^k e^{-\lambda_2}}{k!}$$

Let $R = R_1 + R_2$ with PGF $G$. To find the distribution of $R$ let us evaluate its PGF.

$$G(z) = G_1(z) G_2(z)$$

Where $G_1$ and $G_2$ are the PGF's for $R_1$ and $R_2$ respectively.

$$\begin{aligned} G(z) &= e^{-\lambda_1+\lambda_1 z} e^{-\lambda_2+\lambda_2 z} \\ &= e^{-(\lambda_1+\lambda_2)+(\lambda_1+\lambda_2)z} \end{aligned}$$

But this is the PGF for a random variable with a Poisson distribution with parameter $\lambda_1 + \lambda_2$. Hence the sum of two Poisson distributions is Poisson.

# 3 Central limit theorem

We saw that the sum of mutually independent Bernoulli distributions gives a binomial distribution. This is related to a very powerful result in probability that is known as the *Central Limit Theorem*. It says that if you take the average of $n$ mutually independent random variables that have the same distribution, then the distribution of the average *tends* to the normal distribution as $n$ gets large. This is quite an amazing result since the the distribution of the average tends to a fixed distribution, namely the normal distribution, regardless of the distribution of the random variables in the average.

The normal distribution $N(\mu, \sigma)$ has two parameters, the expectation, $\mu$, and the *standard deviation*, $\sigma$ (which is a measure of how much a random variable deviates from its expectation). Its distribution is given by

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

We won't prove the central limit theorem here (and it won't be on the exam) but it's a neat fact that you'll see when you take a probability course.

# 4 Team Problems

1. This is a real problem that came from a research project in computational biology a few weeks ago. The CS PhD student hadn't taken 6.042 and so he needs our help. Can you solve the problem?

   It is known that atoms occur in different isotopes. For example, Carbon atoms usually have 6 protons and 6 neutrons, but occasionally they have 7 neutrons. In fact 1.11% of all Carbon atoms have 7 neutrons and 98.89% have 6. From what we know, the appearance of the different isotopes seems to be a random event, unrelated to the functionality of the atom. We can tell which Carbon atoms have an extra neutron only by their atomic weight. The atomic weight of the generic Carbon atom is 12 and of its isotope is 13.

The research team is studying proteins composed of Carbon, Oxygen, Nitrogen and Hydrogen (the four building blocks of life). Their atomic weights and isotopes are listed below. The second column is the atomic weight and the third column is the frequency of the isotope.

| | | |
|---|---|---|
| *Hydrogen* | 1 | $P_{H,1}$ |
| | 2 | $P_{H,2}$ |
| *Carbon* | 12 | $P_{C,12}$ |
| | 13 | $P_{C,13}$ |
| *Nitrogen* | 14 | $P_{N,14}$ |
| | 15 | $P_{N,15}$ |
| *Oxygen* | 15 | $P_{O,15}$ |
| | 16 | $P_{O,16}$ |
| | 17 | $P_{O,17}$ |

The research team wants to write software that can predict the distribution of atomic weights for any protein. Describe a simple algorithm for computing the weight distribution of any protein that contains $n_C$ carbon atoms, $n_O$ oxygen atoms, $n_N$ nitrogen atoms and $n_H$ hydrogen atoms. [Hint: Use generating functions!]

2. Coupon Collector problem: We are trying to collect baseball cards. Assume there are $n$ different cards. We can buy them one at a time. When we buy a card, it is equally likely that we get any one of the $n$ cards, and purchases are mutually independent. We want to study how many cards we need to buy before we get one each of the $n$ different kinds.

   (a) Suppose we already have $k$ different cards. What is the expected number of cards we need to buy to get a new card?

   (b) Using the above, what is the expected number of cards we need to buy to get one of each kind?

   (c) Suppose we buy $n \ln n + cn$ cards, where $c > 0$ is a real number. What is the probability that we don't have the $1^{st}$ card?

   (d) Suppose we buy $n \ln n + cn$ cards, what is an upper bound on the probability that we don't have all $n$ cards?