# EE 126 Probability and Random Processes
# January 25, 2018 Poisson Properties, Entropy

## 1 Lecture

$$var(X_1 + X_2 + ... + X_n) = \sum_{i=1}^{n} var(X_i) + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} cov(X_i, X_j)$$

where $cov(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$.

The **correlation coefficient** is defined as

$$\rho(X, Y) = \frac{cov(X, Y)}{\sqrt{var(X)}\sqrt{var(Y)}}$$

Note that $-1 \leq \rho \leq 1$.

*Example: seats on a plane.* Same story: passengers sit in random seats. $X$ is the number of passengers who sit in their actual assigned seats. From last time, we know that $\mathbb{E}[X] = 1$. What is $var(X)$? Again,

$$X_i = \begin{cases} 1 & \text{if } i\text{th passenger is in his assigned seat} \\ 0 & \text{otherwise} \end{cases}$$

and $X = \sum_{i=1}^{n} X_i$. (Each $X_i$ is an **indicator variable**, which means it takes on a value of either 1 or 0. Indicator variables and Bernoulli random variables are the same.)

We have $\mathbb{E}[X_i] = \frac{1}{n}$, and $var(X) = \frac{1}{n}(1 - \frac{1}{n}) = \frac{n-1}{n^2}$. Next, we'll compute the covariance of each $(X_i, X_j)$ pair.

$$cov(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]$$

What is $\mathbb{E}[X_i X_j]$? If we refer to $X_i X_j$ as $z_{ij}$, where

$$z_{ij} = \begin{cases} 1 & \text{if both } i \text{ and } j \text{ are in their seats} \\ 0 & \text{otherwise} \end{cases}$$

then $\mathbb{E}[z_{ij}] = P(z_{ij} = 1) = \frac{1}{n(n-1)}$. And we can put it all together:

$$\begin{aligned} var(X) &= \sum_{i=1}^{n} var(X_i) + \sum_{i} \sum_{j, j\neq i} cov(X_i, X_j) \\ &= \sum_{i=1}^{n} \frac{n-1}{n^2} + \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left( \frac{1}{n(n-1)} - \frac{1}{n^2} \right) \\ &= n\frac{n-1}{n^2} + n(n-1)\left( \frac{1}{n(n-1)} - \frac{1}{n^2} \right) \\ &= \frac{n-1}{n} + \left( 1 - \frac{n-1}{n} \right) \\ &= 1 \end{aligned}$$

Meanwhile, $P(\text{nobody sits in assigned seat}) = \left(\frac{n-1}{n}\right)^n \to \frac{1}{e}$. This is a simplification in which passengers choose seats uniformly at random and can sit on each other's laps.

## Review of Popular Discrete Distributions

- Uniform : $\mathbb{E}[X] = \frac{n-1}{2},\ var(X) = \frac{n^2-1}{12}$
- Bernoulli($p$) : $\mathbb{E}[X] = p,\ var(X) = p(1-p)$
- Binomial($n, p$) : $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k},\ \mathbb{E}[X] = np,\ var(X) = np(1-p)$
- Poisson($\lambda$) : $P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}\ \ (k = 0, 1, ...),\ \mathbb{E}[X] = \lambda,\ var(X) = \lambda$
- Geometric($p$) : $P(X = k) = (1-p)^{k-1} p\ \ (k = 1, 2, ...),\ \mathbb{E}[X] = \frac{1}{p},\ var(X) = ?$

## Memoryless Property - Variance Calculation

Let's calculate $var(X)$, for $X$ a geometric random variable. To do so, we first define the memoryless property of Geom($p$), which states that $P(X = m + n \mid X > m) = P(X = n)$.

As a consequence, $\mathbb{E}[g(X) \mid X > 1] = \mathbb{E}[g(1 + X)]$. For example, $\mathbb{E}[X \mid X > 1] = \mathbb{E}[1 + X] = 1 + \mathbb{E}[X]$.

*Proof.*

$$\begin{aligned}
\mathbb{E}[g(X) \mid X > 1] &= \sum_{k=1}^{\infty} g(k) P(X = k \mid X > 1) \\
&= \sum_{k=1}^{\infty} g(k) P(X = k - 1) \quad \text{(this follows from the memoryless property)} \\
&= \sum_{l=1}^{\infty} g(1 + l) P(X = l) \quad \text{(here, we let } l = k - 1) \\
&= \mathbb{E}[g(1 + X)] \\
&= 1
\end{aligned}$$

Also,

$$\begin{aligned}
\mathbb{E}[X^2] &= \mathbb{E}[X^2 \mid X = 1] P(X = 1) + \mathbb{E}[X^2 \mid X > 1] P(X > 1) \\
&= (1)(p) + \mathbb{E}[(1 + X)^2](1 - p) \\
&= 1 + 2\mathbb{E}[X] + \mathbb{E}[X^2] \\
&= \frac{2 - p}{p^2}
\end{aligned}$$

Thus $var(X) = \frac{2-p}{p^2} - \frac{1}{p^2} = \frac{1-p}{p^2}$.

## Poisson Properties

### Sum of Independent Poissons

Say we have $X \sim \text{Pois}(\lambda),\ Y \sim \text{Pois}(\mu),\ X, Y$ independent. Now, if $Z = X + Y$ then $Z \sim \text{Pois}(\lambda + \mu)$. The sum of independent Poissons is itself Poisson! The intuition for this is based on binomial limiting behavior.

We define

- $X_1 \sim B(n, p_1)$, where $p_1 = \lambda_1 / n$ ($n$ large, $\lambda_1$ constant)
- $X_2 \sim B(n, p_2)$, where $p_2 = \lambda_2 / n$ ($n$ large, $\lambda_2$ constant)

What is a good approximation to $Y = X_1 + X_2$, assuming $X_1$ and $X_2$ are independent? Drawing each $X_i$ as a possible bitstring, we see

$$\begin{aligned} X_1&: 0\ 0\ 0\ 0\ 1\ 0\ 0\ 0\ ...\ 1\ ... \quad (1 \text{ appears with probability } p_1) \\ X_2&: 0\ 0\ 1\ 0\ 0\ 0\ 0\ 0\ ...\ 1\ ... \quad (1 \text{ appears with probability } p_2) \\ Y&: 0\ 0\ 1\ 0\ 1\ 0\ 0\ 0\ ...\ 2\ ... \end{aligned}$$

In the case of $Y$,

- 1 appears with probability $p_1(1-p_2) + p_2(1-p_1) = p_1 + p_2 - 2p_1p_2$
- 2 appears with probability $p_1p_2$
- 0 appears with probability $(1-p_1)(1-p_2) = 1 - p_1 - p_2 - p_1p_2$

*Intuition*: if $p_1 = \lambda_1/n$ and $p_2 = \lambda_2/n$, then $p_1p_2 = \lambda_1, \lambda_2/n^2$. Therefore, for large $n$, 2 will essentially never appear! Let's define events.

$$\begin{aligned} T&: \text{every } Y_i \text{ is 0 or 1 for } i = 1, 2, ..., n \\ A&: \text{at least one } Y_i \text{ is 2 for } i = 1, 2, ..., n \end{aligned}$$

It follows that

$$P(Y = k) = P(Y = k \mid T)P(T) + P(Y = k \mid A)P(A)$$

where

$$P(A) = P\left(\bigcup_{i=1}^{n}(Y_i = 2)\right) \leq \sum_{i=1}^{n} P(Y_i = 2) = \sum_{i=1}^{n} \frac{\lambda_1\lambda_2}{n^2}$$

$$P(A) \leq \frac{\lambda_1\lambda_2}{n}$$

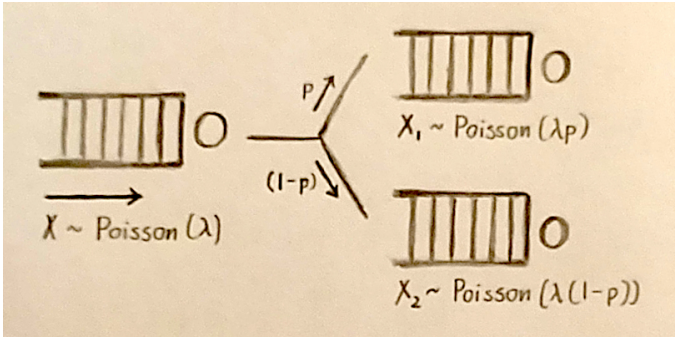
Thus,

$$P(Y = k) = P(Y = k \mid T)\left(1 - \frac{\lambda_1\lambda_2}{n}\right) + P(Y = k \mid A)\left(\leq \frac{\lambda_1\lambda_2}{n}\right)$$

As $n$ goes to $\infty$, $(1 - (\lambda_1\lambda_2)/n)$ goes to 1 and $(\leq (\lambda_1\lambda_2)/n)$ goes to 0. Therefore, at large values of $n$ we have $P(Y = k) = P(Y = k \mid T)$. In other words, we can assume that $Y$ consists only of 0s and 1s (in turn implying that $X_1$ and $X_2$ are never 1 simultaneously). Thus we can model $Y$ as a sum of Bernoulli variables and a Poisson itself.

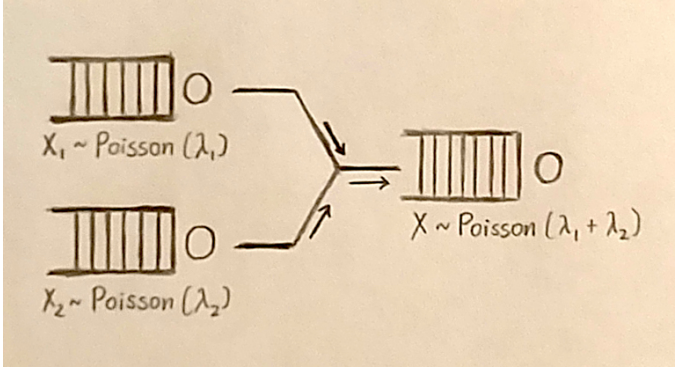
Since only one of $X_{1i}$ and $X_{2i}$ will be 1 for any given $i$, we can simply represent $Y$'s $\lambda$ parameter as $\lambda_1 + \lambda_2$.

**Poisson Splitting**




**Poisson splitting. Each box-like entity represents a queue; each circle represents a server. Packets come in at the rate given by $X$. At the junction, packets go up the top path with probability $p$ and down the bottom path with probability $(1-p)$. We can see that the weights for the individual distributions are split accordingly.**

**Poisson Merging**




This is just another name for a sum of independent Poissons. For a [more] formal proof of Poisson merging, see Discussion 2 of EE 126's Spring 2018 offering.

*From that discussion*: the Poisson distribution is used to model rare events (e.g. the number of customers who enter a store in the next hour). This modeling assumption is justified by the limiting behavior of the binomial distribution, specifically as the number of trials $n$ goes to $\infty$ and the probability of success per trial $p$ goes to 0.

Say we have two independent streams of rare events (e.g. the number of female customers and male customers entering a store). If we do not care to distinguish between the two types of rare events, we can represent the combined stream of events as a single Poisson distribution. This is known as **Poisson merging**.

### Probability and Information: Entropy

Given $P(X = x)$, we have information about the relative frequencies of the values of $X$. Can this information be expressed in a more "fundamental" way?

Suppose there are $N$ possible outcomes of $X$. Then we'll need about $\log_2 N$ bits to represent them (e.g. if $N = 4$, we need to enumerate the four possibilities A, B, C, and D which can be encoded as 00, 01, 10, and 11 respectively). But if $A$ occurs much more frequently than the others, shouldn't we try to use fewer bits on $A$? (Imagine we're sending this information over a network, and there's a price for every bit we send).

We can think of $\log_2 \frac{1}{P(X=k)}$ as being the "surprise" in obtaining $X = k$. (Side note: this was observed by Claude Shannon, the father of information theory.) As a consequence, under our pay-per-bit model we should spend very few bits on things that we expect to happen. If we live in a sunny state, we don't want to spend the same amount of money describing if it's sunny or rainy! Essentially, if an event is really a surprise, we should be fine paying more for it.

Shannon referred to this concept as the "self-information" of $X$.

The expected value of the "surprise" is equal to

$$H(X) = \mathbb{E}\left[\log_2 \frac{1}{P(X = x)}\right] = \sum_x P(X = x) \log_2 \frac{1}{P(X = x)}$$

bits. Note that $H(X)$ represents the minimum number of bits needed to represent a file.

## 2 Extra

If $X_i$ is an indicator variable, then $\mathbb{E}[X_i^2] = \mathbb{E}[X_i]$. We only care about the case where $X_i = 1$.