COMPSCI 70 Self-Study Lecture Notes

Dun-Ming Huang

Contents

Chapter 0

Geometric and Poisson Distributions

0.1 Geometric Distribution

The geometric distribution frequently occurs in situations where we consider "how long we have to wait" for something to occur, such as "how many run until error", or "how many rolls until SSR".

0.1.1 Definition

The **geometric distribution** is then mathematically defined as below:

Definition 0.1.1. Geometric Distribution

A random variable X such that:

$$\mathbb{P}[X = i] = (1-p)^{i-1} p, i \in \mathbb{Z}^+$$

is said to have a geometric distribution with parameter p, mathematically expressable as:

$$X \sim \text{Geometric}(p)$$

Does such distribution sum to 1? Allow us to verify:

$$\begin{aligned}
\sum_{i=1}^{\infty} \mathbb{P}[X = i] &= \sum_{i=1}^{\infty} (1-p)^{i-1} p \\
&= p \sum_{i=1}^{\infty} (1-p)^{i-1} \\
&= p \lim_{n \to \infty} \frac{1-(1-p)^n}{1-(1-p)} \\
&= p \frac{1}{1-1+p} = 1
\end{aligned}$$

The distribution of X would thus be some curve that decreases by a factor of $(1-p)$ at each step; in other words, an exponential decay.

From this fact, we would know that the height of curve is dependent on p, and the speed at which the curve decrease depends on $(1-p)$.

0.1.2 Mean and Variance of a Geometric RV

Let us first discuss the expectation of a geometric random variable.
If directly computed, we would obtain some result like what follows:

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} i\mathbb{P}[X = i] = p\sum_{i=1}^{\infty} i(1-p)^{i-1}$$

However, there is a way to simplify this above complication:

Theorem 0.1.1. Tail Sum Formula

Theorem:

$$\mathbb{E}[X] = \sum_{i=1}^{\infty} \mathbb{P}[X \geq i]$$

Proof:

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i=1}^{\infty} i\mathbb{P}[X = i] \\
&= p\sum_{i=1}^{\infty} i(1-p)^{i-1} \\
&= 0p_0 + 1p_1 + 2p_2 + \cdots \\
&= (p_1 + p_2 + \cdots) + (p_2 + \cdots) + \cdots \\
&= \mathbb{P}[X \geq 1] + \mathbb{P}[X \geq 2] + \cdots
\end{aligned}$$

What is the significance of this?

Theorem 0.1.2. The Expectation of a Geometric RV

$$\begin{aligned}
\mathbb{P}[X \geq i] &= (1-p)(1-p)^{i-1} + p(1-p)^{i-1} \\
&= (1-p)^{i-1} \\
\mathbb{E}[X] &= \sum_{i=1}^{\infty} \mathbb{P}[X \geq i] \\
&= \sum_{i=1}^{\infty} (1-p)^{i-1} = \frac{1}{1-(1-p)} = \frac{1}{p}
\end{aligned}$$

Upon that, let's also discuss and derive such identity:

Theorem 0.1.3. Expectation of X(X - 1)

Theorem:

$$\mathbb{E}[X(X-1)] = \frac{2(1-p)}{p^2}$$

Proof: Preliminarily,

$$\begin{aligned}\mathbb{E}[X(X-1)] &= \sum_{i=1}^{\infty} i(i-1) \times \mathbb{P}[X=i] \\ &= \sum_{i=1}^{\infty} i(i-1) \times p(1-p)^{i-1} \\ &= p(1-p)\sum_{i=1}^{\infty} i(i-1)(1-p)^{i-2}\end{aligned}$$

With some calculus and an identity we just learned,

$$\begin{aligned}\sum_{i=1}^{\infty}(1-p)^i &= \frac{1}{p} \\ \frac{d}{di} &= \sum_{i=1}^{\infty} -i(1-p)^{i-1} \\ &= -\frac{1}{p^2} \\ \frac{d^2}{di^2} &= \sum_{i=1}^{\infty} i(i-1)(1-p)^{i-2} \\ &= \frac{2}{p^3}\end{aligned}$$

Using the result above:

$$\begin{aligned}\mathbb{E}[X(X-1)] &= p(1-p)\sum_{i=1}^{\infty} i(i-1)(1-p)^{i-2} \\ &= p(1-p)\frac{2}{p^3} \\ &= \frac{2(1-p)}{p^2}\end{aligned}$$

Using the linearity of expectation, we can elegantly calculate the variance of a geometric distribution:

Theorem 0.1.4. The Variance of a Geometric RV

Theorem:

$$Var(X) = \frac{1-p}{p^2}$$

Proof:

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

0.1.3 Memoryless

What? Geometric Distributions can have Alzeihmers? This is the first time I have ever seen someone who's not Alzeihmerus be memoryless!
What?

Definition 0.1.2. Memoryless

In memoryless distributions, prior trials do not affect the number of additional trials needed. Mathematically:

$$\mathbb{P}[X > n+m | X > n] = \mathbb{P}[X > n]$$

For example, say we have already tossed the coin m times, what is the probability that we need more than n additional tosses before getting our first head?

$$\begin{aligned}\mathbb{P}[X > n+m | X > m] &= \frac{\mathbb{P}[X > n+m]}{\mathbb{P}[X > m]} \\ &= \frac{(1-p)^{n+m}}{(1-p)^m} \\ &= (1-p)^n = \mathbb{P}[X > n]\end{aligned}$$

In this sense, "memorylessness" refers to how the distribution, when calculating a tail, will not remember the given condition ($X > m$) since it doesn't contribute to the end result of a tail even given that condition.

0.1.4 Example (Application)

What other greater application can CS70 be made on other than probabilistic problems, and more specifically, gacha games!?
Let us look at the "Coupon Collector Problem":

> Let us attempt to collect a set of n different cards.
> We can get one card by performing one summon, each of the n cards equivalently likely to appear from that summon.
> How many summons do we have to do until we have collected at least one copy of every card?

Let S_n denote the number of summoning we need to perform to collect all n cards, then the distribution of S_n is difficult to compute directly. But, its expected value can be calculated via the gift of linearity.
Let X_i be a random variable rerpesenting the number of boxes needed while trying to get the i^{th} new card, starting immediately after the most recent new card's obtain.
Then,

$$S_n = X_1 + X_2 + \cdots + X_n$$

Therefore,

$$\mathbb{E}[S_n] = \mathbb{E}[X_1 + X_2 + \cdots + X_n] = \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_n]$$

The expected value of X_1 would be 1, since any first card is a new card.
Now, how about X_i?
Each time we perform a summoning, we will get an new card with probability $\frac{n-i}{n}$. We will continue performing summoning, which can be characterized as a coin flip of heads probability $p = \frac{n-i}{n}$ for simplicity. This makes X_i the number of tosses until first head.
There,

$$X_i \sim \text{Geometric}(\frac{n-i}{n})$$

And consequentially,

$$\mathbb{E}[X_i] = \frac{n}{n-i}$$

Coming back to the original simplification of $\mathbb{E}[S_n]$, then:

$$\mathbb{E}[S_n] = \sum_{i=1}^{n} \frac{n}{n-i} = n \sum_{i=1}^{n} \frac{1}{i}$$

This can be well approximated as

$$n \sum_{i=1}^{n} \frac{1}{i} \simeq n(\ln n + \gamma_E)$$

Where $\gamma_E = 0.5772\dots$ is known as Euler's constant.

0.2 Poisson Distribution

Lorem Ipsum

0.2.1 Definition

Lorem Ipsum

0.2.2 Mean and Variance of a Poisson RV

Lorem Ipsum

0.2.3 Sum of Independent Poisson RV

Lorem Ipsum

0.2.4 Poisson and Binomial Distributions

Lorem Ipsum