SI140 Discussion 07

Li Zeng, Tao Huang, Xinyi Liu

ShanghaiTech University, China
{zengli,huangtao1,liuxy10}@shanghaitech.edu.cn

1 Normal RVs

We can say that the single most important random variable type is the Normal (or Gaussian) random variable. it originates from the summation of independent random variables and as a result it occurs often in nature. Many things in the world are not distributed normally but data scientists and computer scientists model them as Normal distributions, because it is theoretically proved that normal r.v.s a safest choice of distribution that we can apply to data with a measured mean and variance (Explained in 1.2). From the perspective of pragmatism, though We will not go into this in this discussion, we can say that the Gaussian variable has properties that usually simplifies the algorithms, as we will see in the following lectures.

1.1 Basics

- PDF $X \sim \mathcal{N}(\mu, \sigma^2)$ is $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$.
- CDF of $X \sim \mathcal{N}(0,1)$ is denoted as $\Phi(x)$.
- Linearity: If X is a Normal such that $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$ and Y is a linear transform of X such that $Y = aX + b$ then Y is also a Normal where $Y \sim \mathcal{N}\left(a\mu + b, a^2\sigma^2\right)$.

1.2 Extensions – Theoretic Motivation of Using Gaussian

What is the theoretical motivation for us to use Gaussian r.v.s? There are mainly two. One is Strong Law of Large Numbers (SLLN)[1], and the other is the Central Limit Theorem (CLT)[2]. SLLN roughly tells us that the average of summation will converge to the expectation of the r.v with probability 1, and CLT ensures that the average will converges to a normal distribution once the number of summed r.v.s is large enough. The theorems make normal r.v.s a safest choice of distribution that we can apply to data with a measured mean and variance. Here we list the theorem as reference. You are not expected to prove them.

Theorem 1 (Kolmogorov's Strong Law of Large Numbers (SLLN)). *Suppose $X, X_1, X_2, \ldots$ are i.i.d and $E(X)$ exists. $S_n = \sum_{i=1}^{n} X_i$. Then*

$$S_n/n \to E(X), \quad a.s.$$

Conversely, if $S_n/n \to \mu$ which is finite, then $\mu = E(X)$

Theorem 2 (Central Limit Theorem (CLT)). *If $X_1, X_2, \ldots, X_n$ are random samples each of size n taken from a population with overall mean μ and finite variance σ^2 and if $\bar{X}$ is the sample mean, the limiting form of the distribution of $Z = \left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}\right)$ as $n \to \infty$, is the standard normal distribution $Z \sim \mathcal{N}(0,1)$.*

We can see that with a mild condition (finite variance), we can expect that the sum over samples acts as a normal r.v. Fig. 1 shows the histogram of simulation of summation over different numbers of i.i.d uniform variables. By eye observation, sum over 8 is already really close to a normal distribution.

Exercise 1 (HuaTsing University's Attendance). HuaTsing University accepts 2480 students and each student has a 68% chance of attending. Let $X = \#$ students who will attend. $X \sim \text{Bin}(2480, 0.68)$. What is $P(X > 1745)$?

[1] See more of SLLN at this pdf.
[2] See more of CLT at wiki.

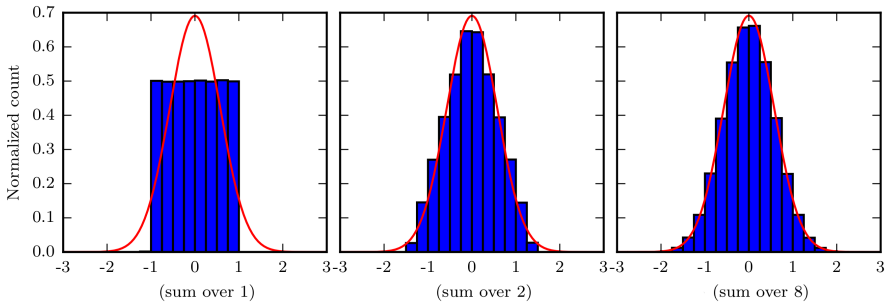


Fig. 1. The histogram of simulation of summation over different numbers of i.i.d uniform r.v.s

Solution 1. $E[X] = np = 1686.4$. $\mathrm{Var}(X) = np(1-p) = 539.7$.$\sigma = \sqrt{\mathrm{Var}(X)} = 23.23$.
We can thus use a Normal approximation: $Y \sim \mathcal{N}(1686.4, 539.7)$

$$P(X > 1745) \approx P(Y > 1745.5) = P\left(\frac{Y - 1686.4}{23.23} > \frac{1745.5 - 1686.4}{23.23}\right) = 1 - \Phi(2.54) = 0.0055$$

.

2 Exponential

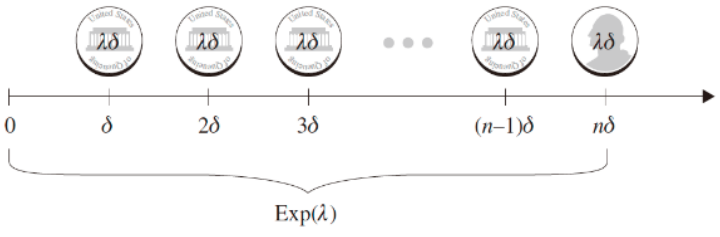
2.1 Basics

- PDF of $X \sim \mathrm{Expo}(\lambda)$ is $f(x) = \lambda e^{-\lambda x}, x > 0$.
- CDF of $X \sim \mathrm{Expo}(\lambda)$ is $F(x) = 1 - e^{-\lambda x}, x > 0$.
- The only memoryless continuous r.v (const failure rate): $P(X \geq s + t \mid X \geq s) = P(X \geq t) \forall s, t > 0$.
- (Expo Clocks' Race) $X_1, \cdots, X_n$ are independent r.v.s, with $X_j \sim \mathrm{Expo}(\lambda_j)$. Let $L = \min(X_1, \cdots, X_n)$. Then:
  - (Time record) $L \sim \mathrm{Expo}(\lambda_1 + \cdots + \lambda_n)$
  - (The winning clock) $P(L = X_k) = \frac{\lambda_k}{\lambda_1 + \cdots + \lambda_n}$

Exercise 2. Prove the two conclusions of the Expo Clocks' Race using basic properties of Exponential r.v.s, then explain intuitively why is the case.

2.2 Extensions

Exercise 3. Explain how we use δ - Steps to relate Geometric distribution with Exponential distribution?



3 Poisson Processes

To consider the need of Poisson Process, let us consider a possible model of traffic accidents within a city. Assuming the traffic intensity to be constant over time, so that the probability rate of accidents should be the same all the time. Under an plausible additional assumption – that different time periods are independent with others – the sequence of accidents happened becomes a Poisson process.

3.1 Definition

Definition 1 (Poisson Process). *An arrival process is called a Poisson process with rate λ if it has the following properties: (a) (Time-homogeneity) The probability $P(k,\tau)$ of k arrivals is the same for all intervals of the same length τ. (b) (Independence) The number of arrivals during a particular interval is independent of the history of arrivals outside this interval. (c) (Small interval probabilities) The probabilities $P(k,\tau)$ satisfy*

$$\begin{aligned} P(0,\tau) &= 1-\lambda\tau+o(\tau) \\ P(1,\tau) &= \lambda\tau+o_1(\tau) \\ P(k,\tau) &= o_k(\tau), \quad \textit{for } k=2,3,\ldots \end{aligned}$$

Here, $o(\tau)$ and $o_k(\tau)$ are functions of τ that satisfy

$$\lim_{\tau\to 0}\frac{o(\tau)}{\tau}=0, \quad \lim_{\tau\to 0}\frac{o_k(\tau)}{\tau}=0$$

No doubt that we see the Bernoulli approximation in (c) in Definition 1. So far we can use the knowledge to connect Poisson distribution with Bernoulli distribution as well as the exponential distribution.

Exercise 4. Explain the following questions:

(a) *(*important) What is the distribution of $Pois(1,\tau)$? How is it approximated by a certain discrete r.v.?
(b) If we want to simulate the traffic accidents by code, how do we do it to make it more easy? Using the original continuous version or using the discrete approximation?

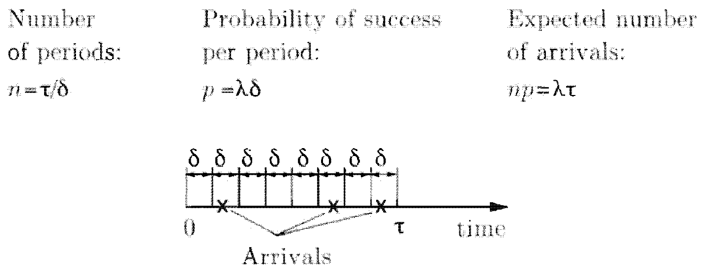


Fig. 2. Bernoulli approximation of the Poisson process over an interval of time

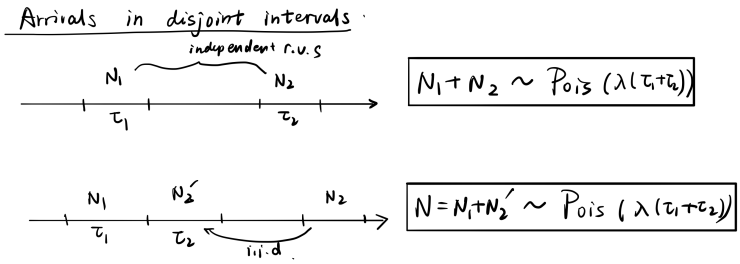


Fig. 3. Arrivals in disjoint intervals in Poisson process.

3.2 Properties

- (Independent Interval) For any given time $t > 0$, the history of the process after time t is also a Poisson process, and is independent from the history of the process until time t.
- (Memoryless) Let t be a given time and let $\bar{T}$ be the time of the first arrival after time t. Then, $\bar{T} - t$ has an exponential distribution with parameter λ, and is independent of the history of the process until time t.
- (Number of Successes on Disjoint Intervals) The number of successes on disjoint intervals can be added and combined. See details in Fig. 3.1.
- (Inter-Arrival Time & Total Time for k Arrivals) Inter-arrival time means the time between two adjacent arrivals, e.g. $T_1, T_1 - T_2, T_3 - T_2, \cdots$ shown in Fig. 3.2. Each of them is a i.i.d Exponential r.v with the same rate parameter with the Poisson process: $t_k = T_k - T_{k-1} \sim Expo(\lambda)$. While the total time for k arrivals T_k follows a Erlang distribution of order k [3], with $E(T_k) = k/\lambda$, $Var(T_k) = k/\lambda^2$.

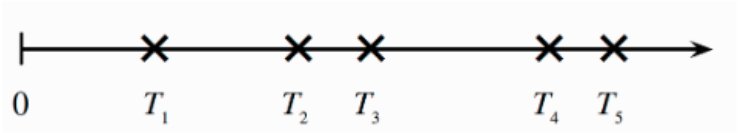


Fig. 4. Inter-arrival Time $t_k = T_k - T_{k-1}$.

- (Splitting and Merging of Poisson Processes) A Poisson process may be splitted into two independent ones if labelling each arrival with a Bernoulli distribution. Two independent Poisson processes can also be merged to be one, adding up the arrivals from two streams. See details in Fig. 3.2.

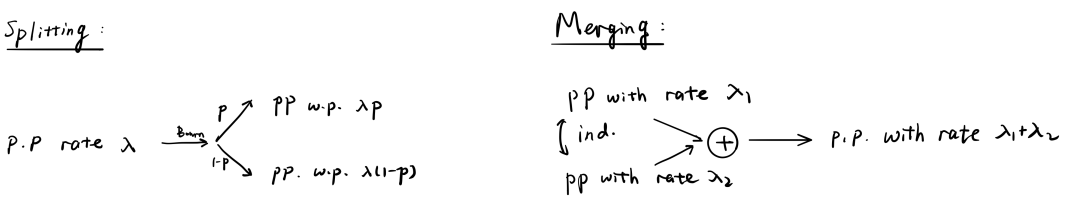


Fig. 5. Splitting and Merging in Poisson process.

Exercise 5. Explain the following questions:

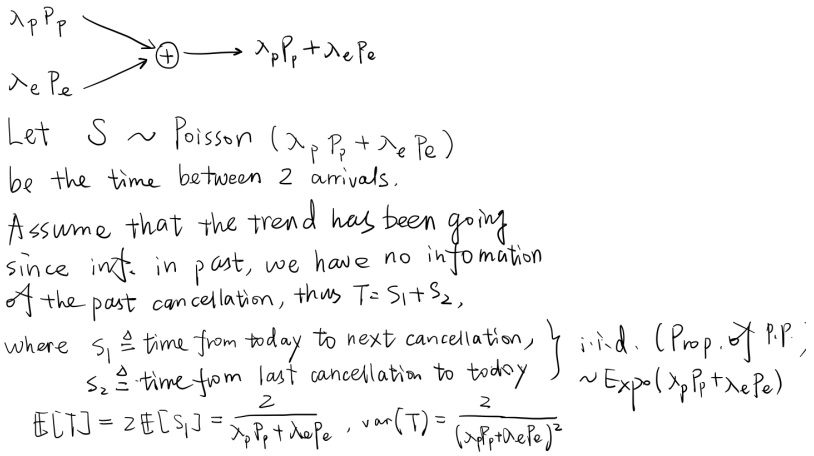
(a) Where does the independence property of Poisson process comes from?
(b) How to get the expectation and variance of total time for k arrivals without knowing Erlang's distribution?
(c) Make justification on the splitting and merging of Poisson processes.
(d) Story-telling on Exponential clock race: Explain the two conclusion derived previously in the context of merging of Poisson processes.
Hint: Perspective from the Bernoulli approximation might be helpful!

Exercise 6 (Last one?). When you enter the bank, you find that all three tellers are busy serving other customers, and there are no other customers in queue. Assume that the service times for you and for each of the customers being served are independent identically distributed exponential random variables. What is the probability that you will be the last to leave?

Solution 2. 1/3.

[3] See more of Erlang distribution here

Exercise 7 (School Cancellation). In the Simpsons' Town, power outages happen according to a Poisson Process with a rate of λ_p and independently earthquakes happen according to a Poisson Process with a rate of λ_e. The headmaster cancels school with probability p_p if there is a power outage, and p_c if there is an earthquake. What is the expectation and the variance of the amount of time T between the previous school cancellation and the next school cancellation from today? You may assume that this trend has been going on since infinitely in the past.



Solution 3.

4 Summary

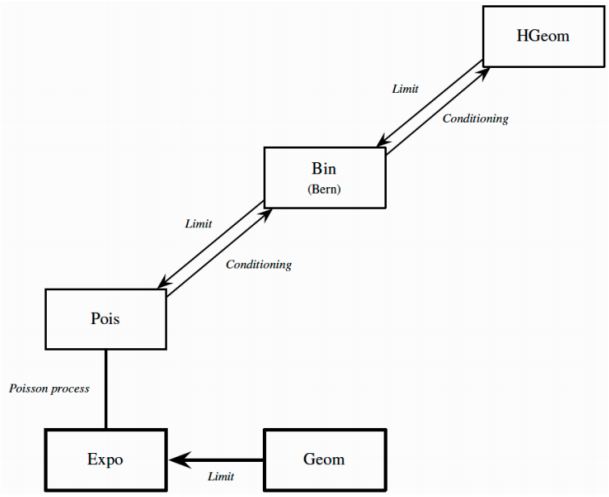


Fig. 6. Relationships between different r.v.s

5 Probabilistic Method

5.1 Main Strategy

Suppose you aim to show that there exists an object in a collection with a certain property, however it is hard (or computationally expensive) to find or construct such an instance, you may as well apply another strategy: *pick an object at random from the collection and show that there is a positive probability of the random object having the desired property.*

5.2 Two Key Ideas

1. The possibility principle: Let A be the event that a randomly chosen object in a collection has a certain property. If $P(A) \geq 0$, then there exists an object with the property.
2. The good score principle: Let X be the score of a randomly chosen object. If $E(X) \geq c$, then there is an object with a score of at least c.

Remark 1. The probabilistic method doesn't tell us how to find an object with the desired property; it only assures us that one exists.

Example 1. (Example from BH with solution, page 187) A group of 100 people are assigned to 15 committees of size 20, such that each person serves on 3 committees. Show that there exist 2 committees that have at least 3 people in common.

Solution 4. A direct approach is inadvisable here: one would have to list all possible committee assignments and compute, for each one, the number of people in common in every pair of committees. The probabilistic method lets us bypass brute-force calculations. To prove the existence of two committees with an overlap of at least three people, we'll calculate the average overlap of two randomly chosen committees in an arbitrary committee assignment. So choose two committees at random, and let X be the number of people on both committees. We can represent $X = I_1 + I_2 + \cdots + I_{100}$ where $I_j = 1$ if the j th person is on both committees and 0 otherwise. By symmetry, all of the indicators have the same expected value, so $E(X) = 100E(I_1)$, and we just need to find $E(I_1)$.

By the fundamental bridge, $E(I_1)$ is the probability that person 1 (whom we'll name Bob) is on both committees (which we'll call A and B). There are a variety of ways to calculate this probability; one way is to think of Bob's committees as 3 tagged elk in a population of 15. Then A and B are a sample of 2 elk, made without replacement. Using the HGeom(3,12,2) PMF, the probability that both of these elk are tagged (i.e., the probability that both committees contain Bob) is $\binom{3}{2}\binom{12}{0}/\binom{15}{2} = 1/35$. Therefore

$$E(X) = 100/35 = 20/7$$

which is just shy of the desired "good score" of 3. But hope is not lost! The good score principle says there exist two committees with an overlap of at least $20/7$, but since the overlap between two committees must be an integer, an overlap of at least $20/7$ implies an overlap of at least 3. Thus, there exist two committees with at least 3 people in common.

Exercise 8 (Two-Coloring Sets). Let each of $A_1, \ldots, A_{500}$ be a set of 10 points on a plane. Some sets may share points. Prove that it is possible to color each point red or blue in such a way that every set A_i has both colors.

Solution 5. The following is a proof by the Probabilistic Method. Color each of the (at most 5000) points independently choosing red or blue with equal likelihood. Then, for $i = 1, \ldots, 500$, the probability that A_i has only one color is $(1/2)^9 = 1/512$. Therefore

$$\begin{aligned} P(\text{there is a set with only one color}) &\leq \sum_{i=1}^{500} P(A_i \text{ has only one color }) \\ &= 500 \times \frac{1}{512} < 1 \end{aligned}$$

It follows from the principle that there exists a coloring without a monochromatic set, that is, one in which every set has both colors. An amazing aspect of this solution lies in that it requires only devising a nice randomization.

Exercise 9 (Ten Points). Prove that any 10 points on a plane can be covered using some number of non-overlapping unit disks (i.e., disks of unit diameter).

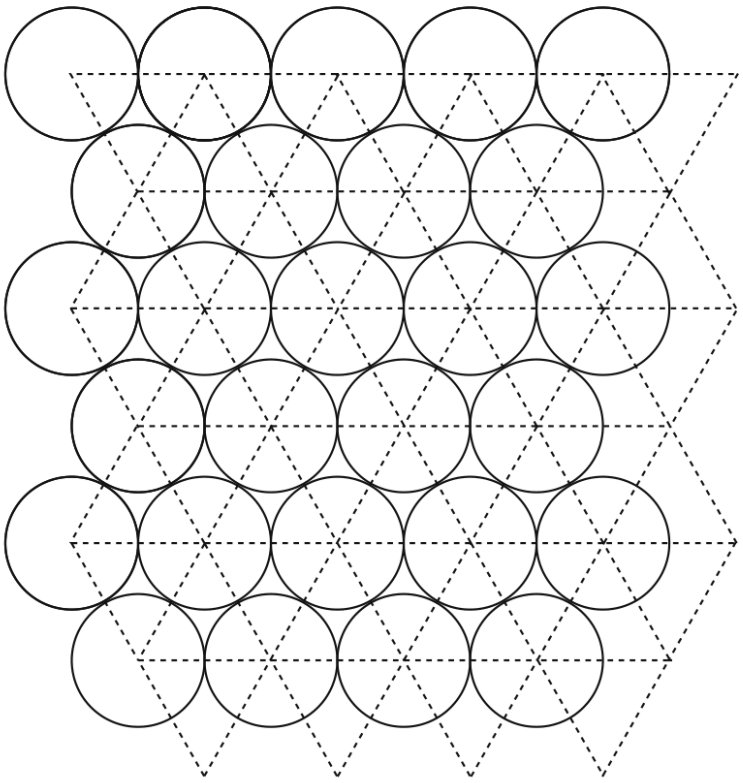


Fig. 7.

Solution 6. Imagine the closest packing of infinitely many disks on the plane. All closest packings look like Figure 7, up to translation and rotation. As is easily calculated, the density of the packing is $\pi/(2\sqrt{3}) \approx 0.9069$. Take such a closest packing randomly. (How to randomize appropriately is explained later.) Then, for any point on the plane, the probability that it is not covered by the chosen packing is about $1 - 0.9069 = 0.0931$. It follows that, for any 10 points $\mathrm{P}_1, \ldots, \mathrm{P}_{10}$,

$$\begin{aligned} P(\text{One or more points are not covered}) &\le \sum_{i=1}^{10} P\left(\mathrm{P}_i \text{ is not covered }\right) \\ &\approx 0.0931 \times 10 = 0.931 < 1 \end{aligned}$$

Therefore, we obtain from the principle that there exists some closest packing that covers all the 10 points. And, in such a packing, we actually need at most 10 disks to cover the 10 points.

The previous proof is written in a rather orthodox way, but the following proof, focusing on expectation, may be easier to understand.

Take a random closest packing. Then, for any point on the plane, the probability that it is covered by the chosen packing is $\pi/(2\sqrt{3}) \approx 0.9069$. It follows that, for any 10 points $\mathrm{P}_1, \ldots, \mathrm{P}_{10}$

$$\begin{aligned}\text{The expected number of covered points} &= \sum_{i=1}^{10} P\left(\mathrm{P}_i \text{ is covered }\right) \\ &\approx 0.9069 \times 10 = 9.069 > 9\end{aligned}$$

That the expected number of covered points is more than 9 implies that there exists the case in which more than 9 points, that is, all the 10 points, are covered, which means that there exists a way to cover all the 10 points.

The Probabilisitic Method was introduced by Paul Erdős, who often talked about *The Book*. According to him, it is a book in which God wrote down the best and most elegant proof for every mathematical theorem. And he always sought proofs from *The Book*. You might agree that the Probabilistic Method is a powerful tool to generate such an elegant proof.