

Milestones of Probability Theory

1 Introduction

Many probabilistic processes can be understood as limits of processes with binomial densities. That is, their densities can be approximated arbitrarily closely by a binomial density $f_{n,p}$ for suitable choices of n and p . In these Notes, we consider two important results of this kind. First, we consider Poisson processes, which are limits of binomial processes where $n \rightarrow \infty$ and $p \rightarrow 0$ while np remains constant. Second, by fixing p and letting n approach infinity, we arrive at a profound result of probability theory: the Central Limit Theorem.

We also consider another fundamental result called the *Strong Law of Large Numbers*. The Strong Law provides important information about how the average of independent trials may vary during the course of the trials. The *Weak Law* we considered in Notes 13-14 is implied by the Strong Law in most circumstances. The Weak Law is also a simple Corollary of the Central Limit Theorem. However, neither the Central Limit Theorem nor the Strong Law imply each other.

2 The Poisson Approximation

We've worked with the binomial distribution, which measures the probability of k successful outcomes occur in a sequence of n independent trials. In this section we'll consider a closely related and widely applicable distribution known as the *Poisson distribution*. The Poisson distribution arises when n is much larger than the expected number of successful outcomes.

2.1 Poisson Random Variables

Let's consider a particular example. Suppose that we want to model the arrival of packets at an internet router. We know that on average the router handles $\lambda = 10^7$ packets per second. Given this expected value, how can we model the actual distribution of packet arrivals in a second? One possible model is to break up each second into tiny intervals of size $\delta > 0$ seconds, so there are a large number, $n = 1/\delta$, of tiny intervals. Then we declare that in each tiny interval, a packet arrives with probability $\lambda\delta$ (this gives the right expected number of arrivals). Under this model, the number, X , of intervals in which a packet actually arrives has a binomial distribution:

$$\Pr \{X = k\} = \binom{1/\delta}{k} (\lambda\delta)^k (1 - \lambda\delta)^{1/\delta - k}. \quad (1)$$

Note that this is not quite the same as counting the number of arrivals, since more than one packet may arrive in a given interval. But if the interval is tiny, this is so unlikely that we can ignore the possibility.

Now we let δ become infinitesimally small (while holding k fixed) and make use of three approximations:

$$\begin{aligned} \binom{1/\delta}{k} &\approx \frac{(1/\delta)^k}{k!} \\ (1 - \lambda\delta)^{1/\delta} &\approx e^{-\lambda} \\ 1 - \delta k &\approx 1. \end{aligned}$$

Plugging these approximations into (1) yields

$$\begin{aligned} \Pr \{X = k\} &= \binom{1/\delta}{k} (\lambda\delta)^k (1 - \lambda\delta)^{1/\delta - k} \\ &= \binom{1/\delta}{k} (\lambda\delta)^k (1 - \lambda\delta)^{(1 - \delta k)/\delta} \\ &\approx \frac{(1/\delta)^k}{k!} (\lambda\delta)^k (1 - \lambda\delta)^{1/\delta} \\ &= \frac{\lambda^k}{k!} (1 - \lambda\delta)^{1/\delta} \\ &\approx \frac{\lambda^k}{k!} e^{-\lambda} \end{aligned} \tag{2}$$

The probability distribution (2) is known as the *Poisson distribution*. When system events appear according to a Poisson density, the system is called a *Poisson process*.

Another example where a Poisson distribution fits the facts is in observing the number of misprints per page in a book. In a well edited book, there may be an average of one misprint on every three pages. That is, there is an average of $\lambda = 1/3$ misprints per page. An average page has about 40 lines of a dozen words, or about $n = 480$ words. If we suppose that each word has an independent probability of $1/(3 \cdot 480)$ of containing an uncorrected misprint, then the density function of errors per page would be $f_{480, 1/1440}$, which will be approximated to three decimal places by the Poisson density with $\lambda = 1/3$.

Further examples of random variables which generally obey a Poisson distribution include:

- the number of decaying particles in a radioactive sample in a given time interval,
- the distribution of the number of failures per day of a system,
- the number of people in a community who are older than 100 years,
- the number of vacancies occurring per year on the Supreme Court,
- the number of wrong telephone numbers dialed in Boston per day.

2.2 Properties of the Poisson Distribution

As a sanity check on our distribution, the probability values (2) had better sum to 1. Using the Taylor expansion for e^λ , we can verify that they do:

$$\sum_{k \in \mathbb{N}} \Pr \{X = k\} = \sum e^{-\lambda} \frac{\lambda^k}{k!} = e^{-\lambda} \sum \frac{\lambda^k}{k!} = e^{-\lambda} e^\lambda = 1.$$

A further sanity check is that the expected number of arrivals in a second is indeed λ , namely,

$$E[X] = \lambda. \quad (3)$$

Similarly, the binomial distribution $f_{n,p}$ has variance $np(1-p)$. Since the Poisson distribution is the limit of $f_{1/\delta, \lambda\delta}$ as δ vanishes, it ought to have variance

$$(1/\delta)(\lambda\delta)(1-\lambda\delta) = \lambda(1-\lambda\delta) \approx \lambda.$$

The final approximation holds since $1-\lambda\delta \approx 1$ for vanishing δ . In other words,

$$\text{Var}[X] = \lambda. \quad (4)$$

Also, suppose we have two independent Poisson processes X_1, X_2 contributing arrival events at the respective rates λ_1, λ_2 . Intuitively, this ought to be the same as having a single process producing independent arrivals at the rate $\lambda_1 + \lambda_2$. This explains another useful property of the Poisson distribution:

Lemma 2.1. *If X_1 and X_2 are Poisson processes, then so is $X_1 + X_2$.*

Both equations (3) and (4), and Lemma 2.1, are easy to verify formally from the definition (2) of the Poisson distribution and the Taylor series for e .

[Optional]

Finally, we can develop a Chernoff style bound for the probability that a Poisson process deviates from its mean. As in the proof of the Chernoff bound, we have for *any* random variable, $R \geq 0$, and constants $c, t \geq 0$, that

$$R \geq c \text{ iff } e^{tR} \geq e^{tc}.$$

So by Markov's inequality

$$\Pr \{R \geq c\} \leq \frac{E[e^{tR}]}{e^{tc}} \quad (5)$$

For a Poisson process, X , we have

$$\begin{aligned} E[e^{tX}] &= \sum_{k \in \mathbb{N}} \frac{e^{tk} e^{-\lambda} \lambda^k}{k!} \\ &= e^{-\lambda} \sum \frac{(\lambda e^t)^k}{k!} \\ &= e^{-\lambda} e^{\lambda e^t} \\ &= e^{\lambda(e^t - 1)}. \end{aligned}$$

So if X is a Poisson process,

$$\Pr \{X \geq c\} \leq e^{\lambda(e^t-1)} e^{-tc} = e^{\lambda(e^t-1)-tc}. \quad (6)$$

To minimize the exponent, we choose $t = \ln(c/\lambda)$, which will be positive as long as $c > \lambda$. Substituting this value for t into (6), we conclude

$$\Pr \{X \geq c\} \leq e^{c \ln(\lambda/c) + c - \lambda}. \quad (7)$$

Now letting, $c = c'\lambda$ in (7) and using (3) yields

$$\begin{aligned} \Pr \{X \geq c' E[X]\} &\leq e^{c'\lambda \ln(1/c') + c'\lambda - \lambda} \\ &= e^{-c'\lambda \ln c' + c'\lambda - \lambda} \\ &= e^{-(c' \ln c' - c' + 1)\lambda} \\ &= e^{-(c' \ln c' - c' + 1) E[X]}. \end{aligned}$$

Notice that this is exactly the same as the Chernoff bound. So we have yet another way in which a Poisson process behaves like a sum of independent indicator variables.

3 The Central Limit Theorem

In the Weak Law of Large Numbers we had $S_n ::= \sum_{i=1}^n G_i$ where G_1, \dots, G_i, \dots were mutually independent variables with same mean, μ , and deviation σ . The Weak Law said that the probability that S_n/n was outside an interval of fixed size $\epsilon > 0$ around μ approached 0 as n approached infinity.

The Central Limit Theorem describes not just the limiting behavior of deviation from the mean of S_n/n , but actually describes a limiting shape of the entire distribution for S_n/n . So this theorem substantially refines the Weak Law.

Definition 3.1. For any random variable R with finite mean, μ_R , and deviation, σ_R , let R^* be the random variable

$$R^* ::= \frac{R - \mu_R}{\sigma_R}.$$

R^* is called the “normalized” version of R .

Note that R^* has mean 0 and deviation 1. In other words, R^* is just R shifted and scaled so that its mean is 0 and its deviation and variance are 1.

The Central Limit Theorem says that *regardless* of the underlying distribution of the variables G_i , so long as they are independent, the distribution of S_n^* converges to the same, *normal*, distribution. It is not surprising that this normal distribution—also known as a *Gaussian* distribution—plays a fundamental role in the study of probability and statistics: as long as you are summing enough random variables, you can pretend that the result is Gaussian.

Definition 3.2. The *normal density function* is the function

$$\eta(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2},$$

and the *normal distribution function* is its integral

$$N(y) = \int_{-\infty}^y \eta(x) dx = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^y e^{-x^2/2} dx.$$

The function $\eta(x)$ defines the standard *Bell curve*, centered about the origin with height $1/\sqrt{2\pi}$ and about two-thirds of its area within unit distance of the origin. The normal distribution function $N(y)$ approaches 0 as $y \rightarrow -\infty$. As y approaches zero from below, $N(y)$ grows rapidly towards $1/2$. Then as y continues to increase beyond zero, $N(y)$ rapidly approaches 1.

Theorem (Central Limit). Let $S_n = \sum_{i=1}^n G_i$ where G_1, \dots, G_i, \dots are mutually independent variables with the same mean, μ , and deviation, σ . Let $\mu_n ::= E[S_n] = n\mu$, and $\sigma_n ::= \sigma_{S_n} = n\sigma$. Now let $S_n^* ::= (S_n - \mu_n)/\sigma_n$ be the normalized version of S_n . Then

$$\lim_{n \rightarrow \infty} \Pr \{S_n^* \leq \beta\} = N(\beta)$$

for any real number β .

To understand the Central Limit Theorem, it helps to see how it implies the Weak Law of Large Numbers.

Note first that $\mu_{S_n} = n\mu$, $\text{Var}[S_n] = n\sigma^2$, and so $\sigma_{S_n} = \sigma\sqrt{n}$. Now,

$$\begin{aligned} \left| \frac{S_n}{n} - \mu \right| > \epsilon & \text{ iff } |S_n - n\mu| > n\epsilon \\ & \text{ iff } \left| \frac{S_n - n\mu}{\sigma_{S_n}} \right| > \frac{n\epsilon}{\sigma_{S_n}} \\ & \text{ iff } |S_n^*| > \frac{\sqrt{n}\epsilon}{\sigma}. \end{aligned}$$

But for any real number $\beta > 0$,

$$\frac{\sqrt{n}\epsilon}{\sigma} > \beta$$

will hold for all large n . Hence, for any $\beta > 0$ and all large n ,

$$\Pr \left\{ \left| \frac{S_n}{n} - \mu \right| > \epsilon \right\} = \Pr \left\{ |S_n^*| > \frac{\sqrt{n}\epsilon}{\sigma} \right\} \leq \Pr \{|S_n^*| > \beta\}. \quad (8)$$

So

$$\begin{aligned} \lim_{n \rightarrow \infty} \Pr \left\{ \left| \frac{S_n}{n} - \mu \right| > \epsilon \right\} & \leq \lim_{n \rightarrow \infty} \Pr \{|S_n^*| > \beta\} && \text{(by (8))} \\ & = \lim_{n \rightarrow \infty} \Pr \{S_n^* > \beta\} + \Pr \{S_n^* < -\beta\} \\ & = 1 - N(\beta) + N(-\beta), && \text{(by the Central Limit Thm (3))} \end{aligned}$$

for all real numbers $\beta > 0$. By choosing β large enough, we can ensure that $N(\beta)$ is arbitrarily close to 1 and $N(-\beta)$ is arbitrarily close to 0, so that final term above is arbitrarily close to $1-1+0 = 0$. Hence,

$$\lim_{n \rightarrow \infty} \Pr \left\{ \left| \frac{S_n}{n} - \mu \right| > \epsilon \right\} = 0,$$

which is the Weak Law of Large Numbers.

We will not prove the Central Limit Theorem, but will only note that a standard proof rests on extending ideas we have already used in deriving Chernoff bounds, in particular properties of $E[e^{tX}]$. Regarded as a function of t , $E[e^{tX}]$ is called the *moment generating function* of the random variable, X . The Central Limit Theorem can be proved using a more complete development of the properties of moment generating functions, more than we have time for in 6.042.

Like the Weak Law of Large Numbers, the Central Limit Theorem as stated cannot be applied to actual problems because the necessary information about the rate of convergence is missing, that is, we need to know the accuracy with which the limit $N(\beta)$ approximates the probability that $S_n^* < \beta$. For variables G_1, G_2, \dots whose absolute value is bounded by about 5, or are themselves normal, a rule of thumb is that (3) holds to one or two decimal places when $|\beta| < 3$ and $n > 30$. But in situations such as those we have seen for designing overload tolerance into systems, and also for ensuring the quality of the solution to an optimization problem by a probabilistic algorithm, we are typically more concerned with events that differ from the mean by many standard deviations. For estimating probabilities at such distribution tails, Chernoff bounds are more accurate than those based on normal distributions. For this reason, Chernoff bounds play a more prominent role in Computer Science than the Central Limit Theorem.

4 Strong Law of Large Numbers [Optional]

[Optional]

We described the *Weak* Law of Large Numbers in previous notes, begging the question of what *strong* law of large numbers we might prove. Roughly speaking, the strong law says that with probability 1, the bound of the weak law will hold for all but a finite number of the S_n *simultaneously*—there will only be finitely many exceptions to it.

Theorem 4.1. [The Strong Law of Large Numbers]¹ Let $S_n ::= \sum_{i=1}^n X_i$ where X_1, \dots, X_i, \dots are mutually independent, identically distributed random variables with finite expectation, μ . Then

$$\Pr \left\{ \lim_{n \rightarrow \infty} \frac{S_n}{n} = \mu \right\} = 1.$$

Although Theorem 4.1 can be proven without this assumption, we will assume for simplicity that the random variables X_i have a finite fourth moment. That is, we will suppose that

$$E[X_i^4] = K < \infty. \quad (9)$$

Proof. To begin, assume that μ , the mean of the X_i , is equal to 0. As usual, let $S_n ::= \sum_{i=1}^n X_i$ and consider

$$E[S_n^4] = E[(X_1 + \dots + X_n) \times (X_1 + \dots + X_n) \times (X_1 + \dots + X_n) \times (X_1 + \dots + X_n)]. \quad (10)$$

Expanding the righthand side of (10) results in terms of the forms

$$X_i^4, \quad X_i^3 X_j, \quad X_i^2 X_j^2, \quad X_i^2 X_j X_k, \quad X_i X_j X_k X_l$$

where i, j, k, l are all different. As all the X_i have mean 0, it follows by independence that

$$\begin{aligned} E[X_i^3 X_j] &= E[X_i^3] E[X_j] = 0 \\ E[X_i^2 X_j X_k] &= E[X_i^2] E[X_j] E[X_k] = 0 \\ E[X_i X_j X_k X_l] &= 0. \end{aligned}$$

¹This section taken from Ross, *A First Course in Probability Theory*.

Now, for a given pair i and j there will be $\binom{4}{2} = 6$ terms in the expansion that will equal $X_i^2 X_j^2$. Hence, after expanding the righthand side (10), we have

$$\begin{aligned} \mathbb{E} [S_n^4] &= n \mathbb{E} [X_i^4] + 6 \binom{n}{2} \mathbb{E} [X_i^2 X_j^2] && \text{(linearity of expectation)} && (11) \\ &= nK + 3n(n-1) \mathbb{E} [X_i^2] \mathbb{E} [X_j^2]. && \text{(by (9) and independence)} && (12) \end{aligned}$$

Now, since

$$0 \leq \text{Var} [X_i^2] = \mathbb{E} [X_i^4] - \mathbb{E}^2 [X_i^2]$$

we see that

$$\mathbb{E}^2 [X_i^2] \leq \mathbb{E} [X_i^4] = K.$$

Therefore, from (12) we have that

$$\mathbb{E} [S_n^4] \leq nK + 3n(n-1)K$$

which implies that

$$\mathbb{E} \left[\frac{S_n^4}{n^4} \right] \leq \frac{K}{n^3} + \frac{3K}{n^2},$$

and so

$$\mathbb{E} \left[\sum_{i=1}^{\infty} \frac{S_n^4}{n^4} \right] = \sum_{i=1}^{\infty} \mathbb{E} \left[\frac{S_n^4}{n^4} \right] \leq K \sum_{i=1}^{\infty} \frac{1}{n^3} + \frac{3}{n^2} < \infty.$$

But since the expected value is finite, the probability that $\sum_{i=1}^{\infty} S_n^4/n^4$ is finite must be one. (If there was a positive probability that the sum is infinite, then its expected value would be infinite.) Now the convergence of a series implies that its n th term goes to 0; so we can conclude that $\lim_{n \rightarrow \infty} S_n^4/n^4 = 0$ with probability 1. But if $S_n^4/n^4 = (S_n/n)^4$ goes to 0, then so must S_n/n ; so we have completed the proof that with probability 1,

$$\frac{S_n}{n} \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

When μ , the mean of the X_i , is not equal to 0, we can apply the preceding argument to the random variables $X_i - \mu$ to obtain that with probability 1,

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{X_i - \mu}{n} = 0$$

or, equivalently,

$$\lim_{n \rightarrow \infty} \sum_{i=1}^n \frac{X_i}{n} = \mu$$

which proves the result. □

We remark that as in the Weak Law, full mutual independence of $\{X_i\}$ is not necessary. The proof above only requires that $\{X_i\}$ are 4-way independent.

4.1 A Failure of the Strong Law

To clarify the somewhat subtle difference between the Weak and Strong Laws of Large Numbers, we will construct an example of a sequence X_1, X_2, \dots of mutually independent random variables that satisfies the Weak Law of Large

Numbers, but not the Strong Law. The distribution of X_i will have to depend on i , because otherwise both laws would be satisfied.²

In particular, let X_1, X_2, \dots be the sequence of mutually independent random variables such that $X_1 = 0$, and for each integer $i > 1$,

$$\Pr\{X_i = i\} = \frac{1}{2i \log i}, \quad \Pr\{X_i = -i\} = \frac{1}{2i \log i}, \quad \Pr\{X_i = 0\} = 1 - \frac{1}{i \log i}.$$

Note that $\mu = E[X_i] = 0$ for all i .

Problem. (a) Show that $\text{Var}[S_n] = \Theta(n^2/\log n)$. *Hint:* $n/\log n > i/\log i$ for $2 \leq i \leq n$.

(b) Show that the sequence X_1, X_2, \dots satisfies the Weak Law of Large Numbers, *i.e.*, prove that for any $\epsilon > 0$

$$\lim_{n \rightarrow \infty} \Pr\left\{\left|\frac{S_n}{n}\right| \geq \epsilon\right\} = 0.$$

We now show that the sequence X_1, X_2, \dots does not satisfy the Strong Law of Large Numbers.

(c) (The first Borel-Cantelli lemma.) Let A_1, A_2, \dots be any infinite sequence of mutually independent events such that

$$\sum_{i=1}^{\infty} \Pr\{A_i\} = \infty. \tag{13}$$

Prove that

$$\Pr\{\text{infinitely many } A_i \text{ occur}\} = 1.$$

Hint: We know that the probability that no A_i with $i \geq r$ occurs is

$$\leq e^{-E[T_r]} \tag{14}$$

where $T_r ::= \sum_{i=r}^{\infty} I_{A_i}$ is the number of events A_i with $i \geq r$ that occur.

(d) Show that $\sum_{i=1}^{\infty} \Pr\{|X_i| \geq i\}$ diverges. *Hint:* $\int dx/(x \log x) = \log \log x$.

(e) Conclude that

$$\Pr\left\{\lim_{n \rightarrow \infty} \frac{S_n}{n} = \mu\right\} = 0. \tag{15}$$

and hence that the Strong Law of Large Numbers *completely* fails for the sequence X_1, X_2, \dots

Hint:

$$\frac{X_n}{n} = \frac{S_n}{n} - \frac{n-1}{n} \frac{S_{n-1}}{n-1},$$

so if $\lim_{n \rightarrow \infty} S_n/n = 0$, then also $\lim_{n \rightarrow \infty} X_n/n = 0$.

²This problem is adapted from Grinstead & Snell, *Intro. to Probability*, Ch.8, exercise 16, pp314–315, where it is credited to David Maslen.