Supervised Learning Project
Course Code: MA 5114

Prepared by: Aditya Khambete
(23B3315)

Supervisor : Prof. Ayan Bhattacharya

Spring 2024-25

Contents

References: Lecture notes of Prof. Ayan Bhattacharya [1] and the book by Lesigne [2]

Chapter 1

Basic Probability

Let (Ω, P) be a finite probability space. [1] Write $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$ and $P(\omega_i) = p_i$. We have the most basic formula of probability i.e.

$$1 = \sum_i p_i, \text{ where each } 0 \leq p_i \leq 1 \tag{1.1}$$

Definition 1.0.1. The probability of an event A is defined as

$$P(A) = \sum_{\omega_i \in A} p_i = \sum_{i=1}^{n} p_i \mathbb{I}_A(\omega_i) \tag{1.2}$$

where $\mathbb{I}_A$ is the indicator function of A. This function maps ω_i to 1 if $\omega_i \in A$ and 0 otherwise.

Some more basic properties of probability are:

- $P(\Omega) = 1$
- $P(A \cup B) = P(A) + P(B)$ if $A \cap B = \emptyset$

Such set Ω is called a sample space, and P is called a probability function. It is easy to see from above properties that

- $P(\emptyset) = 0$
- $P(A^c) = 1 - P(A)$
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Definition 1.0.2. The Probability Space is called uniform if p_i is the same for all ω_i.

1.1 Sequential Experiments

Let us demonstrate this through an elementary example. Assume a binary experiment, taking outcomes 0,1 with q,p respectively. Easy to see that $p + q = 1$. Now, consider we repeat this experiment n times. The sample space in this case is $\Omega_n = \{0, 1\}^n$. The probability of a sequence $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$ is

$$P((\omega_1, \omega_2, \ldots, \omega_n)) = p^{\sum_i \omega_i} q^{n - \sum_i \omega_i} \tag{1.3}$$

This is the probability function defined on the sample space Ω_n. This is a simple example of a product probability space. We say the space $\Omega_n = \{0, 1\}^n$ is equipped with the probability function $P_n = (q, p)^{\otimes n}$, where q,p are the probabilities of 0,1 respectively.
More details on why we did this product come from the notion of independence.

[1] The book doesn't mention the sigma field $\mathscr{F}$.

1.2 Random Variables

Definition 1.2.1. A random variable is a function $X : \Omega \to \mathbb{R}$.

We use the denotion $(X = x)$ for the set $\{\omega \in \Omega : X(\omega) = x\}$. The probability of this event is $P(X = x)$, this is also known as the probability mass function of X. Similarly cumulative distribution function is defined as $F(x) = P(X \leq x)$.

A very underrated fact is the space of random variables is a vector space. This is because the sum of two random variables is also a random variable, so is the product of a random variable with a scalar. The basis of this vector space is the indicator functions of the form $\mathbb{I}_{\omega_i}$ where $\omega_i \in \Omega$. [2]

Definition 1.2.2. Expectation of a random variable X is defined as (if the sum converges)

$$E[X] = \sum_{i=1}^{k} x_i P(X = x_i), \text{ where } k \text{ is the number of distinct values of X} \tag{1.4}$$

Some easy to see properties[3] of expectation are:

- $|E[X]| \leq E[|X|]$
- $E[X] \geq 0$ if $X \geq 0$
- $E[c] = c$ for any constant c, particularly $E[E[X]] = E[X]$
- $E[aX] = aE[X]$
- $E[X + Y] = E[X] + E[Y]$

Say we write $X = \sum_{i=1}^{k} x_i \mathbb{I}_{A_i}$, where $A_i = (X = x_i)$. Then $E[X] = \sum_{i=1}^{k} x_i P(A_i)$, say we take some function $g : \mathbb{R} \to \mathbb{R}$, then write $Y = g(X) = \sum_{i=1}^{k} g(x_i)\mathbb{I}_{A_i}$, hence applying E on both sides, we get

$$E[Y] = E[g(X)] = \sum_{i=1}^{k} g(x_i)P(A_i) \tag{1.5}$$

1.3 Some nice inequalities

Theorem 1.3.1. ***Markov's Inequality:*** *Let X be a non-negative random variable, then for any $a > 0$, we have*

$$P(X \geq a) \leq \frac{E[X]}{a} \tag{1.6}$$

This inequality right above is in some sense the mother of all inequalities. The proof is fairly easy, just use the fact that $P(X \geq a)$ can be written as a summation of $P(X = x_i)$ for $x_i \geq a$, multiply by $\frac{x_i}{a}$ and sum over all x_i.

Theorem 1.3.2. ***Chebyshev's Inequality:*** *Let X be a random variable with finite expectation and variance, then for any $a > 0$, we have*

$$P(|X - E[X]| \geq a) \leq \frac{Var[X]}{a^2} \tag{1.7}$$

Follows from Markov's inequality, just use the fact that $Var[X] = E[(X - E[X])^2]$, and apply Markov's inequality on $Y = (X - E[X])^2$.

Definition 1.3.3. The variance of a random variable X is defined as

$$Var[X] = E[(X - E[X])^2], \text{ it simplifies to } E[X^2] - E[X]^2 \tag{1.8}$$

[2] I_{ω_i} is the function that maps ω_i to 1 and all other ω_j to 0. As one might expect, this is a random variable as well.
[3] The last 2 facts imply E is a linear functional on the vector space of random variables.

Chapter 2

Independence

Definition 2.0.1. Two events A, B are independent if $P(A \cap B) = P(A)P(B)$. Except the trivial case with $P(B) = 0$, we can write this as $\frac{P(A \cap B)}{P(B)} = \frac{P(A)}{P(\Omega)}$.

Once we have the notion of independence, we can check where does the product probability space come from.

$$P((\omega_1, \omega_2, \ldots, \omega_n)) = p^{\sum_i \omega_i} q^{n - \sum_i \omega_i} \tag{2.1}$$

Let us extend the notion to more than 1 event,

Definition 2.0.2. A set of events $A_1, A_2, \ldots, A_n$ are independent if for any subset $I \subset \{1, 2, \ldots, n\}$, we have

$$P(\cap_{i \in I} A_i) = \prod_{i \in I} P(A_i) \tag{2.2}$$

One very important thing to note is that if the events $A_1, A_2, \ldots, A_n$ satisfy $P(\cap_{i \in I} A_i) = \prod_{i \in I} P(A_i)$ it doesn't imply that the events are independent, just take one of the events to be the empty set for example. An interesting example showing that pairwise independence doesn't imply mutual independence is the following. Say we havre a fair coin, define,

$$A_1 = \{HH, HT\}$$
$$A_2 = \{HH, TH\}$$
$$A_3 = \{HT, TH\}$$

Then A_1, A_2, A_3 are pairwise independent, but not mutually independent.

2.1 Independence of Random Variables

Random variables are independent if the events $\{X_i = x\}$ are independent for all $x \in$

al.$Alternatively, we can say that the events (\mathrm{X}_1 \in B_1)$, $(X_2 \in B_2)$, ..., $(X_n \in B_n)$ are independent for all $B_1, B_2, \ldots, B_n \subset \mathbb{R}$. Its not hard to notice the following facts:

- independence of random variables doesn't depend on the order of the random variables.
- random variables $X_1, X_2, \ldots, X_n$ are independent if the events ($X_1 = x_1$ and $X_2 = x_2$ and ... and $X_{j-1} = x_{j-1}$)[1] is independent of $(X_j = x_j)$ for all $j \in \{2, 3, \ldots, n\}$, $x_1, x_2, \ldots, x_{j-1}, x_j \in \mathbb{R}$.
- 2 events are independent if and only if the indicator functions are independent.

Now notice the following fact,

Theorem 2.1.1. *If X, Y are independent random variables, then*

$$E[XY] = E[X]E[Y] \tag{2.3}$$

[1] Writing ($X_1 = x_1$ and $X_2 = x_2$) is the same as writing $(X_1 = x_1) \cap (X_2 = x_2)$

Proof. Let A and B be the range of X, Y respectively,
We have

$$\begin{aligned} E[XY] &= \sum_{x\in A}\sum_{y\in B} xyP(X=x, Y=y) \\ &= \sum_{x\in A}\sum_{y\in B} xyP(X=x)P(Y=y) \\ &= \sum_{x\in A} xP(X=x)\sum_{y\in B} yP(Y=y) && \{\text{By Independence}\} \\ &= E[X]E[Y] \end{aligned}$$

□

Corollary 2.1.2. *If X, Y are independent random variables, then $Var[X+Y] = Var[X] + Var[Y]$*

Proof.

$$\begin{aligned} Var[X+Y] &= E[(X+Y)^2] - E[X+Y]^2 \\ &= E[X^2] + E[Y^2] + 2E[XY] - E[X]^2 - E[Y]^2 - 2E[X]E[Y] \\ &= Var[X] + Var[Y] \end{aligned}$$

□

Chapter 3

Binomial Distribution

Take the product probability space $\Omega_n = \{0,1\}^n$ with $P_n = (q,p)^{\otimes n}$. Let S_n be the number of 1's in the sequence $\omega = (\omega_1, \omega_2, \ldots, \omega_n)$. Then S_n is a random variable on Ω_n.

Theorem 3.0.1. *The probability mass function of S_n is given by*

$$P(S_n = k) = \binom{n}{k} p^k q^{n-k} \tag{3.1}$$

We say the random variable S_n follows a binomial distribution with parameters n, p. Note that q is not a parameter, it is just $1-p$.
To prove this, just notice the probability of selecting a specific sequence with k 1's is $p^k q^{n-k}$, and the number of such sequences is $\binom{n}{k}$.

Definition 3.0.2. Bernoulli distribution is a special case of binomial distribution with $n = 1$. It is the same as a unbiased coin flip. Few properties-

- $E[X] = p$
- $Var[X] = pq = p(1-p)$

Theorem 3.0.3. *Let $X_1, X_2, \ldots, X_n$ be independent Bernoulli random variables with parameter p. Then the sum $S_n = X_1 + X_2 + \ldots + X_n$ follows a binomial distribution with parameters $n, p.s$*

The proof is fairly easy from the results we have already proved.

Theorem 3.0.4. *$E[S_n] = np$ and $Var[S_n] = npq$*

Proof.

$$\begin{aligned} E[S_n] &= E[X_1 + X_2 + \ldots + X_n] \\ &= E[X_1] + E[X_2] + \ldots + E[X_n] \\ &= np \end{aligned}$$

For variance, we have

$$\begin{aligned} Var[S_n] &= Var[X_1 + X_2 + \ldots + X_n] \\ &= Var[X_1] + Var[X_2] + \ldots + Var[X_n] \\ &= npq \end{aligned}$$

□

Chapter 4

Kolmogorov's Consistency Theorem

4.1 Introduction

We will now shift our discussion to understanding limits of random variable, which allows us to move towards the limit theorems such as Law of Large Numbers, Central Limit Theorem etc.

The fundamental concept of limits of random variable require us to understand what does sequence of Random Variables mean, and does it even exist. This needs a powerful theorem in measure theory called Kolmogorov's Consistency (Extension) Theorem.

Roughly speaking, Kolmogorov's Consistency Theorem states that if we have some 'consistent' collection of distribution, then there exists a schocastic process that has those distributions.

4.2 Formal Framework

To rigorously frame this, let's define few terms.

Definition 4.2.1 (Schocastic Process)**.** Let $\mathbb{X}$ a collection of random variables $\{X_t\}_{t\in T}$, where T is an index set, such that each X_t is a random variable on a probability space $(\Omega, \mathcal{F}, P)$, and range of X_t is $\mathcal{R}$. Then $\mathbb{X}$ is called a schocastic process if $\{\mathbb{X}(\omega) = (X_t(\omega) : t \in T) \in B\} \in \mathcal{F}$ for all $B \in \mathcal{B}(\mathcal{R}^T)$.

For the above definition, let us define what we mean by finite dimensional distributions.

Definition 4.2.2 (Finite Dimensional Distributions)**.** Let $\mathbb{X}$ be a schocastic process. Then the finite dimensional distributions of $\mathbb{X}$ are the distributions of the random vectors $(X_{t_1}, X_{t_2}, \ldots, X_{t_n})$ for all $n \in \mathbb{N}$ and $t_1, t_2, \ldots, t_n \in T$.

Note that when we say 'distribution' we mean that of a random vector, so it is nothing but a multivariate function which should satisfy the properties of a typical cumulative distribution function.

Definition 4.2.3 (Indistinguishability of Schocastic Processes)**.** Two schocastic processes $\mathbb{X}$ and $\mathbb{Y}$ are said to be indistinguishable if they have the same finite dimensional distributions. Denoted by $\mathbb{X} \overset{d}{=} \mathbb{Y}$.

Definition 4.2.4 (Consistent Collection of Distributions)**.** A collection of distributions $\mathscr{D}$ is said to be consistent if it satisfies the following properties:

1. Each function $\mathbb{F}_{t_1,t_2,\ldots,t_k}$ in $\mathscr{D}$ is a distribution function.
2. $\mathbb{F}_{t_1,t_2,\ldots,t_k}(x_1, x_2, \ldots, x_{k-1}, \infty) = \mathbb{F}_{t_1,t_2,\ldots,t_{k-1}}(x_1, x_2, \ldots, x_{k-1})$ for all $k \geq 2$.
3. For a permutation π of $\{1, 2, \ldots, k\}$, $\mathbb{F}_{t_1,t_2,\ldots,t_k}(x_1, x_2, \ldots, x_k) = \mathbb{F}_{t_{\pi(1)},t_{\pi(2)},\ldots,t_{\pi(k)}}(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(k)})$.

Now we are ready to state Kolmogorov's Consistency Theorem.

Theorem 4.2.5. *Let $\mathbb{P} = \{P_n\}_{n\in\mathbb{N}}$ be a consistent collection of distributions. Then there exists a schocastic process $\mathbb{X}$, and a probability space $(\Omega, \mathcal{F}, P)$ with $P(\bigcap_{i=1}^{k}\{X_{t_i} \leq x_i\}) = \mathbb{F}_{t_1,t_2,\ldots,t_k}(x_1, x_2, \ldots, x_k)$ for every $(x_1, x_2, \ldots, x_k) \in \mathbb{R}^k$ and $k \geq 1$. Moreover, $\mathbb{X}$ is unique in distribution.*

Now we have the theorem for general schocastic process, lets limit ourselves to index set $T = \mathbb{N}$, which is nothing but a sequence of random variables. Before that, let's quickly define what is 'consistency' for this particular case.

Definition 4.2.6. A collection of distributions $\mathscr{L}$ is said to be consistent if it satisfies the following properties:

1. Each function $\mathbb{F}_{1,2,3\ldots k}$ in $\mathscr{L}$ is a distribution function.
2. $\mathbb{F}_{1,2,3\ldots k}(x_1, x_2, \ldots, x_{k-1}, \infty) = \mathbb{F}_{1,2,3\ldots k-1}(x_1, x_2, \ldots, x_{k-1})$ for all $k \geq 2$.

An interesting thing to note is how the above definition is equivalent to the Definition 2.4, where T is just $\mathbb{N}$. That is easy to see, since every permutation of a finite subset of $\mathbb{N}$ has a canonical order which is just the ascending order.

Corollary 4.2.7. *Let $\mathscr{L}$ be a consistent collection of distributions. Then there exists a sequence of random variables $\mathbb{X}$ and a probability space such that $\mathscr{L}$ is the finite dimensional distributions of $\mathbb{X}$.*

Chapter 5

Weak Law of Large Numbers

The Kolmogorov's Theorem gives us the existence of a sequence of random variables, so now once we have that, let the sequence be $\{X_n\}_{n\in\mathbb{N}}$, where X_n are independent and identically distributed random variables. For now, let's assume the 2nd moment of X_n exists (hence the variance exists). We are interested in the sample mean of the first n random variables, which is given by $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Now let's say $E(X_1) = \mu$ (hence $E(X_k) = \mu$ for all $k \leq n$)

$$\begin{aligned} E(\bar{X}_n) &= E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\ &= \frac{1}{n}\sum_{i=1}^{n} E(X_i) \\ &= \frac{1}{n}\sum_{i=1}^{n} \mu \\ &= \mu \end{aligned}$$

So we have $E(\bar{X}_n) = \mu$ for all n. Now let's calculate the variance of $\bar{X}_n$.

$$\begin{aligned} Var\left(\sum_{i=1}^{n} X_i\right) &= \sum_{i=1}^{n} Var(X_i) + 2\sum_{1\leq i<j\leq n} Cov(X_i, X_j) \\ &= nVar(X_1) \quad \text{(Since } X_i \text{ are i.i.d., hence covariances vanish)} \end{aligned}$$

Hence, we get

$$Var(\bar{X}_n) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i) = n\frac{1}{n^2}Var(X_1) = \frac{1}{n}Var(X_1)$$

Now we know variance is a measure of dispersion, and as $n \to \infty$, $Var(\bar{X}_n) = \frac{1}{n}Var(X_1) \to 0$. Which means that the sample mean $\bar{X}_n$ converges to the true mean μ as $n \to \infty$.

To state it formally, let us define L_p convergence and convergence in probability.

Definition 5.0.1 (L_p Convergence)**.** Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables and X be a random variable. Then $\{X_n\}_{n\in\mathbb{N}}$ is said to converge in L_p to X if $E(|X_n - X|^p) \to 0$ as $n \to \infty$.

Definition 5.0.2 (Convergence in Probability)**.** Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables and X be a random variable. Then $\{X_n\}_{n\in\mathbb{N}}$ is said to converge in probability to X if for all $\epsilon > 0$, $P(|X_n - X| > \epsilon) \to 0$ as $n \to \infty$.

It is easy to see that L_p convergence for any $p > 0$ implies convergence in probability. Notice, since the variance of $\bar{X}_n$ goes to 0 as $n \to \infty$, $\bar{X}_n$ converges in L_2 to μ, hence it converges in probability to μ. This is the Weak Law of Large Numbers, let us state it formally.

Theorem 5.0.3 ((A weaker version [1] of) Weak Law of Large Numbers)**.** *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of i.i.d. random variables with $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$. Then the sample mean $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ converges in probability to μ.*

[1] We are nowhere near SLLN yet, as it says about almost sure convergence

The above theorem is a weaker version of the Weak Law of Large Numbers, since we assume variance (and hence 2nd moment) exists. The actual theorem does not require the variance to exist, and is a stronger result, now notice the idea of the proof of the above theorem, we used the fact that variance goes to 0 as $n \to \infty$, this idea can be generalized for any $p > 1$, but now we need some way to say it for $p = 1$, which only assume the existence of the first moment, which indeed is the weak law of large numbers. But the catch is, it is not easy to do just that, and we need some more notions to do that, which we will not mention here.

The experimental probability of success is $\frac{S_n}{n}$, intutively we expect this to be close to p as n increases. This is the weak law of large numbers.

Theorem 5.0.4. *Let $X_1, X_2, \ldots, X_n$ be independent and identically distributed (iid) random variables with mean μ. Let $S_n = X_1 + X_2 + \ldots + X_n$. Then for any $\epsilon > 0$, we have*

$$P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) \to 0 \text{ as } n \to \infty \tag{5.1}$$

Proof. Use Chebyshev's inequality, we have

$$\begin{aligned} P\left(\left|\frac{S_n}{n} - \mu\right| \geq \epsilon\right) &\leq \frac{Var[S_n/n]}{\epsilon^2} \\ &= \frac{Var[S_n]}{n^2\epsilon^2} \\ &= \frac{n\sigma^2}{n^2\epsilon^2} && \text{(By Independence)} \\ &= \frac{\sigma^2}{n\epsilon^2} \\ &\to 0 \text{ as } n \to \infty \end{aligned}$$

□

This has a nice application in analysis, the problem of uniformly approximating a continuous function by a polynomial. This is known as Weierstrass Approximation Theorem. To state it rigorously, let $f : [0,1]^2 \to \mathbb{R}$ be a continuous function. Then for any $\epsilon > 0$, there exists a polynomial $P(x)$ such that $|f(x) - P(x)| < \epsilon$ for all $x \in [0,1]$.
Serge Bernstein gave a probabilistic proof of this theorem.

Lemma 5.0.5. *Let f be a continuous function on $[0,1]$. Then,*

$$\sup_{x\in[0,1]} |f(x) - P_n(x)| \to 0 \text{ as } n \to \infty \text{ ,where } P_n(x) = \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k}x^k(1-x)^{n-k} \tag{5.2}$$

Proof. Fix $\epsilon > 0$, $\exists\ \eta$ such that

$$|x - y| < \eta \implies |f(x) - f(y)| < \epsilon \text{ where } 0 \leq x, y \leq 1 \tag{5.3}$$

Consider the space (Ω_n, P_n), and the random variable $f\left(\frac{S_n}{n}\right)$ where S_n is as above. We have

$$\begin{aligned} E\left[f\left(\frac{S_n}{n}\right)\right] &= \sum_{k=0}^{n} f\left(\frac{k}{n}\right) P(S_n = k) \\ &= \sum_{k=0}^{n} f\left(\frac{k}{n}\right)\binom{n}{k}p^k(1-p)^{n-k} \end{aligned}$$

By WLLN, we have

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \eta\right) \to 0 \text{ as } n \to \infty \tag{5.4}$$

[2]This can be generalised to any closed interval [a,b]

Hence, $\exists\ N_0$ independent of p, such that for every $n \geq N_0$

$$P\left(\left|\frac{S_n}{n} - p\right| > \eta\right) < \epsilon \tag{5.5}$$

Also we have

$$\left|E_n\left[f\left(\frac{S_n}{n}\right)\right] - f(p)\right| = \left|\sum_{k=0}^{n}\left(f\left(\frac{k}{n}\right) - f(p)\right)P_n(S_n = k)\right|$$

breaking the sum into 2 parts, and applying the triangle inequality, we get the following upper bound,

$$\begin{aligned}
&\sum_{\left|\frac{k}{n}-p\right|\leq\eta}\left|f\left(\frac{k}{n}\right) - f(p)\right|P_n(S_n = k) + \sum_{\left|\frac{k}{n}-p\right|>\eta}\left(\left|f\left(\frac{k}{n}\right)\right| + |f(p)|\right)P_n(S_n = k)\\
\leq &\sum_{\left|\frac{k}{n}-p\right|\leq\eta}\epsilon P_n(S_n = k) + \sum_{\left|\frac{k}{n}-p\right|>\eta} 2\sup_{0\leq x\leq 1}|f(x)|P_n(S_n = k)\\
= &\epsilon + 2\sup_{0\leq x\leq 1}|f(x)|P_n\left(\left|\frac{S_n}{n} - p\right| > \eta\right)\\
= &\epsilon + 2\sup_{0\leq x\leq 1}|f(x)|\epsilon
\end{aligned}$$

Which shows the upper bound can be made arbitrarily small, hence the proof. □

Chapter 6

Large Deviations Estimate

As we saw in the last chapter, the WLLN tells us the sample mean $\bar{X}_n$ of i.i.d. random variables converges in probability to the true mean μ. For instance, in n independent Bernoulli trials with success probability p, the proportion of successes $\frac{S_n}{n}$ converges to p. But notice, we still don't know anything about the rate of this convergence. For many applications, knowing this is crucial. Going forward, let's shift our focus to $X_1 \sim Ber(p)$ again.

The Chebyshev inequality, which we used to prove a version of the WLLN, provides a bound:

$$P\left(\left|\frac{S_n}{n} - p\right| \geq \epsilon\right) \leq \frac{Var(S_n/n)}{\epsilon^2} = \frac{p(1-p)}{n\epsilon^2}. \tag{6.1}$$

This bound gives us an idea that the probability decreases as $O(\frac{1}{n})$. However, it turns out, the actual probability in the above equation, decays much faster, exponentially in p. This is the base of large-deviations theory.

6.1 The Large Deviations Estimate for Bernoulli Trials

We will now present the result for the Bernoulli case, which shows that the rate of convergence is indeed exponential. Let S_n be as usual.

First, for every $\epsilon \in (0, 1-p)$, we define the function:

$$h_+(\epsilon) := (p+\epsilon)\ln\frac{p+\epsilon}{p} + (1-p-\epsilon)\ln\frac{1-p-\epsilon}{1-p}.$$

This function will appear in the exponent of our bound. It can be shown that $h_+(\epsilon) > 0$ for $\epsilon \in (0, 1-p)$.

Theorem 6.1.1 (Large Deviations Estimate for Bernoulli Upper Tail)**.** *For every $\epsilon \in (0, 1-p)$ and $n \geq 1$, we have*

$$P\left(\frac{S_n}{n} \geq p+\epsilon\right) \leq e^{-nh_+(\epsilon)}.$$

A similar bound holds for the lower tail, $P(\frac{S_n}{n} \leq p-\epsilon)$, by defining $h_-(\epsilon) := h_+(-\epsilon)$ for $0 < \epsilon < p$.

Proof of Theorem 6.1.1. [1] Fix $t > 0$. Since e^x is an increasing function, the event $\frac{S_n}{n} \geq p+\epsilon$ is equivalent to $S_n \geq n(p+\epsilon)$, which in turn is equivalent to $tS_n \geq tn(p+\epsilon)$. Thus,

$$P\left(\frac{S_n}{n} \geq p+\epsilon\right) = P(S_n \geq n(p+\epsilon)) = P(e^{tS_n} \geq e^{tn(p+\epsilon)}).$$

By Markov's inequality, since $Y = e^{tS_n}$ is non-negative, we get

$$P(e^{tS_n} \geq e^{tn(p+\epsilon)}) \leq \frac{E[e^{tS_n}]}{e^{tn(p+\epsilon)}}.$$

Now, $S_n = \sum_{i=1}^n X_i$ where X_i are i.i.d. Bernoulli(p) random variables.

[1] The proof uses a technique known as Chernoff's bounding method.

Now as we can see, in the numerator we have a term for $E(e^{tS_n})$ we want to calculate this, for some $t \in \mathbb{R}, t > 0$, so we start with

$$E[e^{tS_n}] = E[e^{t(X_1+X_2+\ldots+X_n)}] = E[e^{tX_1}e^{tX_2}\cdots e^{tX_n}].$$

Because the random variables X_i are independent, the expectation of their product is the product of their expectations:

$$E[e^{tX_1}e^{tX_2}\cdots e^{tX_n}] = E[e^{tX_1}]E[e^{tX_2}]\cdots E[e^{tX_n}].$$

Since all X_i are identically distributed, $E[e^{tX_i}]$ is the same for all i. Let's calculate this for a single X_i:

$$\begin{aligned} E[e^{tX_i}] &= \sum_{k\in\{0,1\}} e^{tk}P(X_i = k) \\ &= e^{t\cdot 0}P(X_i = 0) + e^{t\cdot 1}P(X_i = 1) \\ &= (1)\cdot(1-p) + e^t \cdot p \\ &= 1 - p + pe^t. \end{aligned}$$

Therefore, substituting this back, we get:

$$E[e^{tS_n}] = (E[e^{tX_1}])^n = (1 - p + pe^t)^n.$$

[2] Substituting this into our inequality:

$$P\left(\frac{S_n}{n} \geq p + \epsilon\right) \leq \frac{(1-p+pe^t)^n}{e^{tn(p+\epsilon)}} = \left((1-p+pe^t)e^{-t(p+\epsilon)}\right)^n.$$

It is cleat that this holds $\forall\ t > 0$. To get the tightest possible bound, we minimize the term inside the parenthesis with respect to t. This is equivalent to minimizing its log, or maximizing the negative of its log:

$$\ln\left((1-p+pe^t)e^{-t(p+\epsilon)}\right) = \ln(1-p+pe^t) - t(p+\epsilon).$$

Let $g(t) = t(p+\epsilon) - \ln(1-p+pe^t)$. We wish to maximize $g(t)$ for $t > 0$. Taking the derivative with respect to t and putting it zero:

$$g'(t) = p + \epsilon - \frac{pe^t}{1-p+pe^t} = 0.$$

Solving for e^{t^*}:

$$\begin{aligned} (p+\epsilon)(1-p+pe^{t^*}) &= pe^{t^*} \\ (p+\epsilon)(1-p) + (p+\epsilon)pe^{t^*} &= pe^{t^*} \\ (p+\epsilon)(1-p) = pe^{t^*} - (p+\epsilon)pe^{t^*} &= pe^{t^*}(1-(p+\epsilon)) = pe^{t^*}(1-p-\epsilon). \end{aligned}$$

So, $e^{t^*} = \frac{(p+\epsilon)(1-p)}{p(1-p-\epsilon)}$. For this t^* to be positive, we need $e^{t^*} > 1$, which means $(p+\epsilon)(1-p) > p(1-p-\epsilon)$. $p - p^2 + \epsilon - \epsilon p > p - p^2 - p\epsilon \implies \epsilon > 0$, which is true by assumption. Also, for t^* to be well-defined, we need $p + \epsilon < 1$ and $1 - p - \epsilon > 0$.

The value of the exponent at this optimal t^* is $-\sup_{t>0} g(t)$. It is easy to see that $\sup_{t>0} g(t)$ is precisely $h_+(\epsilon)$. Thus, $P\left(\frac{S_n}{n} \geq p + \epsilon\right) \leq e^{-nh_+(\epsilon)}$. □

A similar argument yields for the lower tail, and the related function $h_-(\epsilon)$

$$P(\frac{S_n}{n} \leq p - \epsilon) \leq e^{-nh_-(\epsilon)}$$

For the particular case of bernouli, it follows from simply interchanging p by $1-p$, and S_n by $n - S_n$

[2] The term $E[e^{tX}]$ is called the moment generating function (mgf) for some random variable X. Turns out for the sum S_n of iid r.v.s $\{X_i\}_{i=1}^n$,$\text{mgf}_{S_n}(t) = (\text{mgf}_{X_1}(t))^n$

6.2 Optimality of the bound

The bound $e^{-nh_+(\epsilon)}$ we obtain here is not just an arbitrary upper bound; it is, in a logarithmic sense, the best possible exponential bound.

Proposition 6.2.1 (Optimality of the Exponent)**.** *For every $\epsilon \in (0, 1-p)$,*

$$\lim_{n\to\infty} \frac{1}{n} \ln P\left(\frac{S_n}{n} \geq p+\epsilon\right) = -h_+(\epsilon).$$

Proof. From Theorem 6.1.1 we have, $P\left(\frac{S_n}{n} \geq p+\epsilon\right) \leq e^{-nh_+(\epsilon)}$. Taking the logarithm, dividing by n, and taking the limit superior, we get:

$$\limsup_{n\to\infty} \frac{1}{n} \ln P\left(\frac{S_n}{n} \geq p+\epsilon\right) \leq -h_+(\epsilon).$$

To complete the proof, we need to show the corresponding lower bound for the limit inferior:

$$\liminf_{n\to\infty} \frac{1}{n} \ln P\left(\frac{S_n}{n} \geq p+\epsilon\right) \geq -h_+(\epsilon).$$

Let $k_n = \lceil n(p+\epsilon) \rceil$. Since $\epsilon \in (0, 1-p)$, we have $p < p+\epsilon < 1$. For large n, k_n is an integer such that $0 < k_n < n$. Also, as $n \to \infty$, $k_n/n \to p+\epsilon$. Let $q_n = k_n/n$. So, $q_n \to p+\epsilon$.

The event $\{S_n/n \geq p+\epsilon\}$ includes the event $\{S_n = k_n\}$ because $k_n/n = \lceil n(p+\epsilon)\rceil/n \geq n(p+\epsilon)/n = p+\epsilon$. Thus,

$$P\left(\frac{S_n}{n} \geq p+\epsilon\right) \geq P(S_n = k_n).$$

The probability $P(S_n = k_n)$ is given by the binomial formula:

$$P(S_n = k_n) = \binom{n}{k_n} p^{k_n}(1-p)^{n-k_n}.$$

We use Stirling's approximation for the factorial[3], which yields approximation for binomial coefficient $\binom{n}{k} \sim \frac{1}{\sqrt{2\pi n(k/n)(1-k/n)}} e^{nH(k/n)}$, where $H(q) = -\ln q - (1-q)\ln(1-q)$. So, for $q_n = k_n/n$:

$$\binom{n}{k_n} \sim \frac{1}{\sqrt{2\pi n q_n(1-q_n)}} e^{nH(q_n)}.$$

Therefore,

$$\begin{aligned} P(S_n = k_n) &\sim \frac{1}{\sqrt{2\pi n q_n(1-q_n)}} e^{nH(q_n)} p^{nq_n}(1-p)^{n(1-q_n)} \\ &= \frac{1}{\sqrt{2\pi n q_n(1-q_n)}} e^{n[-q_n \ln q_n - (1-q_n)\ln(1-q_n) + q_n \ln p + (1-q_n)\ln(1-p)]} \\ &= \frac{1}{\sqrt{2\pi n q_n(1-q_n)}} e^{-n[q_n \ln(q_n/p) + (1-q_n)\ln((1-q_n)/(1-p))]}. \end{aligned}$$

Let $f(q) = q\ln(q/p) + (1-q)\ln((1-q)/(1-p))$. As $n \to \infty$, $q_n = k_n/n \to p+\epsilon$. Since $f(q)$ is continuous for $q \in (0,1)$,

$$\lim_{n\to\infty} \left[q_n \ln \frac{q_n}{p} + (1-q_n)\ln\frac{1-q_n}{1-p}\right] = (p+\epsilon)\ln\frac{p+\epsilon}{p} + (1-(p+\epsilon))\ln\frac{1-(p+\epsilon)}{1-p} = h_+(\epsilon).$$

Now consider $\frac{1}{n}\ln P(S_n = k_n)$. Since $A_n \sim B_n$ implies $\ln A_n = \ln B_n + o(1)$,

$$\begin{aligned} \frac{1}{n}\ln P(S_n = k_n) &= \frac{1}{n}\ln\left(\frac{1}{\sqrt{2\pi n q_n(1-q_n)}}\right) - \left[q_n \ln\frac{q_n}{p} + (1-q_n)\ln\frac{1-q_n}{1-p}\right] + \frac{o(1)}{n} \\ &= -\frac{\ln(2\pi n q_n(1-q_n))}{2n} - \left[q_n \ln\frac{q_n}{p} + (1-q_n)\ln\frac{1-q_n}{1-p}\right] + o(1/n). \end{aligned}$$

As $n \to \infty$:

[3]Proved in next chapter

- $q_n \to p + \epsilon$. Since $p + \epsilon \in (0, 1)$, $q_n(1 - q_n)$ converges to $(p + \epsilon)(1 - (p + \epsilon))$, which is a positive constant.
- Thus, $\ln(2\pi n q_n(1 - q_n)) \sim \ln(Cn)$ for some constant $C > 0$.
- So, $\lim_{n\to\infty} -\frac{\ln(2\pi n q_n(1-q_n))}{2n} = 0$.
- The term in square brackets converges to $h_+(\epsilon)$.
- The $o(1/n)$ term vanishes.

Therefore,

$$\lim_{n\to\infty} \frac{1}{n} \ln P(S_n = k_n) = -h_+(\epsilon).$$

Since $P\left(\frac{S_n}{n} \geq p + \epsilon\right) \geq P(S_n = k_n)$, we have

$$\ln P\left(\frac{S_n}{n} \geq p + \epsilon\right) \geq \ln P(S_n = k_n).$$

Dividing by n and taking the limit inf:

$$\liminf_{n\to\infty} \frac{1}{n} \ln P\left(\frac{S_n}{n} \geq p + \epsilon\right) \geq \lim_{n\to\infty} \frac{1}{n} \ln P(S_n = k_n) = -h_+(\epsilon).$$

Combining the lim sup result and this lim inf result, we conclude:

$$\lim_{n\to\infty} \frac{1}{n} \ln P\left(\frac{S_n}{n} \geq p + \epsilon\right) = -h_+(\epsilon).$$

□

6.3 Extension to other random variables

The method used in the 6.1.1 is quite generalizable, and can be extended to sums $S_n = \sum_{i=1}^n X_i$ of other independent and identically distributed (i.i.d.) random variables, provided the moment generating function (MGF) exists. Let $X_1, X_2, \ldots$ be i.i.d. random variables with mean $E[X_1] = \mu$. Let $M_X(t) = E[e^{tX_1}]$ be their common MGF, assumed to exist for t in some interval around 0. Let $K_X(t) = \ln M_X(t)$ be the cumulant generating function (CGF).

Following the same steps as in the Bernoulli proof:
1. For $x > \mu$ (e.g., $x = \mu + \epsilon$), and $t > 0$:

$$P\left(\frac{S_n}{n} \geq x\right) = P(S_n \geq nx) \leq \frac{E[e^{tS_n}]}{e^{tnx}} = \frac{(M_X(t))^n}{e^{tnx}} = \left(M_X(t)e^{-tx}\right)^n.$$

2. To find the tightest bound, we minimize $M_X(t)e^{-tx}$ with respect to $t > 0$, which is equivalent to maximizing $tx - K_X(t)$:

$$P\left(\frac{S_n}{n} \geq x\right) \leq e^{-n \sup_{t>0}[tx - K_X(t)]}.$$

The function $I(x) = \sup_t[tx - K_X(t)]$ (where the supremum is taken over t such that $K_X(t)$ is defined, and typically $t > 0$ if $x > \mu$, $t < 0$ if $x < \mu$) is called the *rate function* Thus, the general bound, called the 'Chernoff Bound' is:

$$P\left(\frac{S_n}{n} \geq x\right) \leq e^{-nI(x)} \quad \text{for } x > \mu.$$

And similarly,

$$P\left(\frac{S_n}{n} \leq x\right) \leq e^{-nI(x)} \quad \text{for } x < \mu.$$

Proposition 6.2.1, which is known as 'Cramer's Theorem' in its general form, states that under certain regularity conditions on X_i, this rate $I(x)$ is indeed optimal in the logarithmic sense:

$$\lim_{n\to\infty} \frac{1}{n} \ln P\left(\frac{S_n}{n} \approx x\right) = -I(x).$$

The function $h_+(\epsilon)$ we saw for the Bernoulli case is simply $I(p+\epsilon)$ when $X_i \sim \text{Bernoulli}(p)$. Each distribution with a well-defined MGF will have its own specific rate function $I(x)$.

This idea forms the foundation of Large Deviation Theory, which allows us to quantify the exponentially small probabilities of such types of events.

Chapter 7

The Central Limit Theorem

7.1 Towards Central Limit Theorem

Now we have the weak law of large numbers, let's try to understand the Central Limit Theorem. The Central Limit Theorem roughly states that the sum of a large number of i.i.d. random variables is approximately normally distributed. Let's try to understand this.

Let us just roughly go through the intuition of the Central Limit Theorem. Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables with $E[X_1] = \mu$ and $Var(X_1) = \sigma^2 > 0$. Consider the sum $S_n = \sum_{i=1}^{n} X_i$. We know $E[S_n] = n\mu$ and $Var(S_n) = n\sigma^2$. The Weak Law of Large Numbers tells us that the sample mean $\bar{X}_n = S_n/n$ converges in probability to μ. This implies that for large n, S_n/n is concentrated around μ, and thus S_n is concentrated around $n\mu$. Now consider the centered sum $S_n - n\mu$. Clearly, $E[S_n - n\mu] = E[S_n] - n\mu = n\mu - n\mu = 0$. The variance is $Var(S_n - n\mu) = Var(S_n) = n\sigma^2$. If we simply look at $S_n - n\mu$, its variance $n\sigma^2 \to \infty$ as $n \to \infty$ (unless $\sigma^2 = 0$, which is a trivial case). If we look at $(S_n - n\mu)/n = \bar{X}_n - \mu$, its variance is $Var(\bar{X}_n - \mu) = Var(\bar{X}_n) = \sigma^2/n \to 0$ as $n \to \infty$. This means $\bar{X}_n - \mu$ converges to a random variable with mean 0 and variance 0, i.e., it converges to the constant 0. This is consistent with the WLLN, but it results in a degenerate distribution in the limit.

To obtain a non-degenerate limiting distribution, we need to find a normalization that keeps the variance stable and positive as $n \to \infty$. Consider the standardized sum:

$$Z_n = \frac{S_n - E[S_n]}{\sqrt{Var(S_n)}} = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

Let's check the mean and variance of Z_n:

$$E[Z_n] = E\left[\frac{S_n - n\mu}{\sigma\sqrt{n}}\right] = \frac{1}{\sigma\sqrt{n}} E[S_n - n\mu] = \frac{1}{\sigma\sqrt{n}} \cdot 0 = 0.$$

$$Var(Z_n) = Var\left(\frac{S_n - n\mu}{\sigma\sqrt{n}}\right) = \frac{1}{(\sigma\sqrt{n})^2} Var(S_n - n\mu) = \frac{1}{n\sigma^2} \cdot n\sigma^2 = 1.$$

So, for every n, Z_n is a random variable with mean 0 and variance 1. This normalization gives us hope that Z_n might converge to some non-degenerate distribution with mean 0 and variance 1. That distribution, remarkably, turns out to be the standard normal distribution $N(0, 1)$. This is the essence of the Central Limit Theorem.

7.2 Statement of the Theorem

The Central Limit Theorem is fascinating because of the extremely wide range of applications, and it establishes the fundamental role of the normal (or Gaussian) distribution, the famous bell curve, it is the reason, of what's so 'normal' about the normal distrubution.

Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed random variables with $E[X_1] = \mu$ and $Var(X_1) = \sigma^2$ Let $S_n = \sum_{i=1}^{n} X_i$. Then $E[S_n] = n\mu$ and $Var(S_n) = n\sigma^2$. The standardized sum is $Z_n = \frac{S_n - n\mu}{\sqrt{n\sigma^2}}$.

Theorem 7.2.1 (Central Limit Theorem). *Let a and b be two elements of $\mathbb{R} \cup \{+\infty\} \cup \{-\infty\}$ such that $a < b$. Then*

$$\lim_{n\to\infty} P\left(a \leq \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2}\, dx.$$

The integral on the right-hand side represents the probability that a standard normal random variable $Z \sim N(0,1)$ falls between a and b. We often denote $\Phi(b) - \Phi(a)$, where $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-x^2/2}\, dx$ is the cumulative distribution function (CDF) of the standard normal distribution.

Note how we didn't comment on the type of convergence in the introductory intuition. The theorem above states convergence of probabilities, which is equivalent to convergence in distribution of the standardized sum Z_n to a standard normal random variable. This means the CDF of Z_n converges pointwise to the CDF of $N(0,1)$ at all the continuity points, which is $\forall t \in \mathbb{R}$. It can be shown that this convergence is, in fact, uniform in both a and b

7.3 Remarks on the Normal Distribution

The function $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is the probability density function (PDF) of the standard normal distribution, often called the Gaussian curve. An important property is that its integral over the entire real line is 1:

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{-x^2/2}\, dx = 1.$$

The integral of $e^{-x^2/2}$ cannot be expressed in terms of elementary functions, but its definite integrals can be computed numerically to really high precision. The function $\Phi(y) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{y} e^{-x^2/2}\, dx$ is extensively tabulated and can be calcualted using any standard statistical software.

7.4 Proof for the Bernoulli Case

We will for this section restrict ourselves again to only sum of iid bernouli. In this simple-looking case aswell, the proof of Theorem 7.2.1 is composed of several steps. The most important instrument is Stirling's formula which approximates $n!$, which allows us to estimate binomial probabilities $P(S_n = k)$. This leads to the de Moivre-Laplace theorem, which gives an approximation for $P(S_n = k)$. The final step involves summing these local probabilities, which approximates a Riemann sum for the integral of the normal density.

Proposition 7.4.1 (Stirling's Formula). *For each integer $n > 0$,*

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e}\right)^n,$$

meaning $\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}(n/e)^n} = 1$. More precisely, $n! = \sqrt{2\pi n} n^n e^{-n}(1 + \epsilon_n)$, where $\epsilon_n = O(1/n)$

Proof. First, we will show that there exists a $c_1 \in \mathbb{R}$ such that

$$\ln(n!) = c_1 + \left(n + \frac{1}{2}\right) \ln n - n + O\left(\frac{1}{n}\right).$$

This estimate is based on a comparison of the series with general term $\ln n$ to the logarithmic integral. We write

$$\ln(n!) = \sum_{k=1}^{n} \ln k = \int_{1/2}^{n+1/2} \ln t\, dt + \sum_{k=1}^{n} \left(\ln k - \int_{k-1/2}^{k+1/2} \ln t\, dt\right). \tag{7.1}$$

On one hand,

$$\int_{1/2}^{n+1/2} \ln t\, dt = [t \ln t - t]_{1/2}^{n+1/2} = \left(n + \frac{1}{2}\right) \ln\left(n + \frac{1}{2}\right) - \left(n + \frac{1}{2}\right) - \left(\frac{1}{2} \ln \frac{1}{2} - \frac{1}{2}\right).$$

Let $c_2 = -(\frac{1}{2}\ln\frac{1}{2} - \frac{1}{2})$. The Taylor series for the logarithm function about 1 yields $\ln(1+x) = x - x^2/2 + O(x^3)$. Thus,

$$\ln\left(n+\frac{1}{2}\right) = \ln\left(n\left(1+\frac{1}{2n}\right)\right) = \ln n + \ln\left(1+\frac{1}{2n}\right) = \ln n + \frac{1}{2n} - \frac{1}{2}\left(\frac{1}{2n}\right)^2 + O\left(\frac{1}{n^3}\right) = \ln n + \frac{1}{2n} + O\left(\frac{1}{n^2}\right).$$

So,

$$\begin{aligned}\left(n+\frac{1}{2}\right)\ln\left(n+\frac{1}{2}\right) &= \left(n+\frac{1}{2}\right)\left(\ln n + \frac{1}{2n} + O\left(\frac{1}{n^2}\right)\right)\\ &= n\ln n + \frac{1}{2} + O\left(\frac{1}{n}\right) + \frac{1}{2}\ln n + O\left(\frac{\ln n}{n}\right)\\ &= \left(n+\frac{1}{2}\right)\ln n + \frac{1}{2} + O\left(\frac{\ln n}{n}\right).\end{aligned}$$

Therefore,

$$\int_{1/2}^{n+1/2} \ln t\,dt = \left(n+\frac{1}{2}\right)\ln n + \frac{1}{2} - n - \frac{1}{2} + c_2 + O\left(\frac{\ln n}{n}\right) = \left(n+\frac{1}{2}\right)\ln n - n + c_3 + O\left(\frac{1}{n}\right), \quad (7.2)$$

for some constant $c_3 = c_2$.

On the other hand, consider the term $\ln k - \int_{k-1/2}^{k+1/2} \ln t\,dt$.

$$\int_{k-1/2}^{k+1/2} \ln t\,dt = [t\ln t - t]_{k-1/2}^{k+1/2} = \left(k+\frac{1}{2}\right)\ln\left(k+\frac{1}{2}\right) - \left(k-\frac{1}{2}\right)\ln\left(k-\frac{1}{2}\right) - 1.$$

Using $\ln(k \pm 1/2) = \ln k + \ln(1 \pm 1/(2k)) = \ln k \pm \frac{1}{2k} - \frac{1}{8k^2} \pm \frac{1}{24k^3} + O(1/k^4)$:

$$\begin{aligned}\left(k+\frac{1}{2}\right)\ln\left(k+\frac{1}{2}\right) &= \left(k+\frac{1}{2}\right)\left(\ln k + \frac{1}{2k} - \frac{1}{8k^2} + O\left(\frac{1}{k^3}\right)\right)\\ &= k\ln k + \frac{1}{2} - \frac{1}{8k} + \frac{1}{2}\ln k + \frac{1}{4k} + O\left(\frac{1}{k^2}\right)\\ &= k\ln k + \frac{1}{2}\ln k + \frac{1}{2} + \frac{1}{8k} + O\left(\frac{1}{k^2}\right).\end{aligned}$$

$$\begin{aligned}\left(k-\frac{1}{2}\right)\ln\left(k-\frac{1}{2}\right) &= \left(k-\frac{1}{2}\right)\left(\ln k - \frac{1}{2k} - \frac{1}{8k^2} + O\left(\frac{1}{k^3}\right)\right)\\ &= k\ln k - \frac{1}{2} - \frac{1}{8k} - \frac{1}{2}\ln k + \frac{1}{4k} + O\left(\frac{1}{k^2}\right)\\ &= k\ln k - \frac{1}{2}\ln k - \frac{1}{2} + \frac{1}{8k} + O\left(\frac{1}{k^2}\right).\end{aligned}$$

Subtracting these:

$$\left(k+\frac{1}{2}\right)\ln\left(k+\frac{1}{2}\right) - \left(k-\frac{1}{2}\right)\ln\left(k-\frac{1}{2}\right) = \ln k + 1 + O\left(\frac{1}{k^2}\right).$$

So we have,

$$\begin{aligned}\int_{k-1/2}^{k+1/2} \ln t\,dt &= [t\ln t - t]_{k-1/2}^{k+1/2}\\ &= \left(k+\frac{1}{2}\right)\ln\left(k+\frac{1}{2}\right) - \left(k-\frac{1}{2}\right)\ln\left(k-\frac{1}{2}\right) - 1\\ &= \ln k + 1 + O\left(\frac{1}{k^2}\right) - 1 = \ln k + O\left(\frac{1}{k^2}\right)\end{aligned}$$

So, we get

$$\ln k - \int_{k-1/2}^{k+1/2} \ln t \, dt = O\left(\frac{1}{k^2}\right).$$

Thus, the series $\sum_{k=1}^{\infty} \left(\ln k - \int_{k-1/2}^{k+1/2} \ln t \, dt\right)$ is absolutely convergent. Let

$$c_5 := \sum_{k=1}^{\infty} \left(\ln k - \int_{k-1/2}^{k+1/2} \ln t \, dt\right).$$

Then,

$$\sum_{k=1}^{n} \left(\ln k - \int_{k-1/2}^{k+1/2} \ln t \, dt\right) = c_5 - \sum_{k=n+1}^{\infty} \left(\ln k - \int_{k-1/2}^{k+1/2} \ln t \, dt\right). \tag{7.3}$$

Since the general term is $O(1/k^2)$, the tail sum $\sum_{k=n+1}^{\infty} O(1/k^2)$ is $O(1/n)$. So, $\sum_{k=1}^{n} \left(\ln k - \int_{k-1/2}^{k+1/2} \ln t \, dt\right) = c_5 + O\left(\frac{1}{n}\right)$.

From (7.1), (7.2), and (7.3):

$$\ln(n!) = \left(n + \frac{1}{2}\right) \ln n - n + c_3 + c_5 + O\left(\frac{1}{n}\right).$$

Let $c_1 = c_3 + c_5$. Then $\ln(n!) = \left(n + \frac{1}{2}\right) \ln n - n + c_1 + O\left(\frac{1}{n}\right)$. Exponentiating, we get $n! = e^{c_1} n^{n+1/2} e^{-n} e^{O(1/n)}$. Since $e^{O(1/n)} = 1 + O(1/n)$,

$$n! = d \cdot n^{n+1/2} e^{-n} (1 + \epsilon_n),$$

where $d = e^{c_1}$ and $\epsilon_n = O(1/n)$.

To complete the proof, we must show that $d = \sqrt{2\pi}$. To do this, we use Wallis' integrals, defined for $m \in \mathbb{N}$ by

$$I_m := \int_0^{\pi/2} \sin^m t \, dt.$$

Integration by parts yields $I_m = \int_0^{\pi/2} \sin^{m-1} t \sin t \, dt$.

$$I_m = [-\sin^{m-1} t \cos t]_0^{\pi/2} - \int_0^{\pi/2} (m-1) \sin^{m-2} t \cos t (-\cos t) \, dt = (m-1) \int_0^{\pi/2} \sin^{m-2} t \cos^2 t \, dt.$$

So, $I_m = (m-1) \int_0^{\pi/2} \sin^{m-2} t (1 - \sin^2 t) \, dt = (m-1)(I_{m-2} - I_m)$.

This yields the recurrence $mI_m = (m-1)I_{m-2}$ for $m \geq 2$. We have $I_0 = \pi/2$ and $I_1 = \int_0^{\pi/2} \sin t \, dt = [-\cos t]_0^{\pi/2} = 1$. For $n \geq 1$:

$$I_{2n} = \frac{2n-1}{2n} \frac{2n-3}{2n-2} \cdots \frac{1}{2} I_0 = \frac{(2n-1)!!}{(2n)!!} \frac{\pi}{2} = \frac{(2n)!}{(2^n n!)^2} \frac{\pi}{2}.$$

$$I_{2n+1} = \frac{2n}{2n+1} \frac{2n-2}{2n-1} \cdots \frac{2}{3} I_1 = \frac{(2n)!!}{(2n+1)!!} = \frac{(2^n n!)^2}{(2n+1)!}.$$

Since $0 \leq \sin t \leq 1$ for $t \in [0, \pi/2]$, $\sin^m t \leq \sin^{m-1} t \leq \sin^{m-2} t$.
So $I_m \leq I_{m-1} \leq I_{m-2}$. Dividing by I_m: $1 \leq I_{m-1}/I_m \leq I_{m-2}/I_m = m/(m-1)$.
As $m \to \infty$, $m/(m-1) \to 1$, so by the sandwich Theorem, $\lim_{m\to\infty} \frac{I_{m-1}}{I_m} = 1$.
And, from the explicit formulas mentioned above, we get:

$$\frac{I_{2n}}{I_{2n+1}} = \frac{(2n)!\pi}{2^{2n}(n!)^2 2} \cdot \frac{(2n+1)!}{2^{2n}(n!)^2} = \frac{((2n)!)^2(2n+1)\pi}{2^{4n}(n!)^4 2}.$$

Now, substitute $n! \sim dn^{n+1/2}e^{-n}$ into this limit expression. $(2n)! \sim d(2n)^{2n+1/2}e^{-2n}$.

$$\begin{aligned}\frac{((2n)!)^2(2n+1)\pi}{2^{4n}(n!)^4 \cdot 2} &\sim \frac{(d(2n)^{2n+1/2}e^{-2n})^2(2n+1)\pi}{2^{4n}(dn^{n+1/2}e^{-n})^4 \cdot 2} \\ &= \frac{d^2(2n)^{4n+1}e^{-4n}(2n+1)\pi}{2^{4n}d^4n^{4n+2}e^{-4n} \cdot 2} \\ &= \frac{d^2 2^{4n+1}n^{4n+1}e^{-4n}(2n+1)\pi}{2^{4n}d^4n^{4n+2}e^{-4n} \cdot 2} \\ &= \frac{d^2 \cdot 2 \cdot n^{4n+1}(2n+1)\pi}{d^4n^{4n+2} \cdot 2} = \frac{(2n+1)\pi}{d^2 n}.\end{aligned}$$

As $n \to \infty$, $\frac{(2n+1)\pi}{d^2 n} \to \frac{2\pi}{d^2}$. Since the limit is 1, we have $\frac{2\pi}{d^2} = 1$, which implies $d^2 = 2\pi$, so $d = \sqrt{2\pi}$ (since d must be positive). This completes the proof. □

Proposition 7.4.2 (de Moivre-Laplace Theorem)**.** *Let S_n be the sum of n i.i.d. Bernoulli(p) random variables. Let k be an integer. Let $x_k = \frac{k-np}{\sqrt{np(1-p)}}$. If x_k remains in a bounded interval as $n \to \infty$ (which means k is not too far from np, specifically $|k - np| = O(\sqrt{n})$), then*

$$P(S_n = k) = \binom{n}{k} p^k (1-p)^{n-k} \sim \frac{1}{\sqrt{2\pi np(1-p)}} e^{-x_k^2/2}$$

as $n \to \infty$. The convergence is uniform for k such that $|k - np| \le a\sqrt{n}$ for any fixed $a > 0$.

Proof Sketch. This theorem is proved by applying Stirling's formula to $n!$, $k!$, and $(n-k)!$ in the binomial coefficient $\binom{n}{k}$. From Stirling's formula, we have for $k \in I_n = [np - a\sqrt{n}, np + a\sqrt{n}]$:

$$\binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}.$$

Applying Stirling's $m! \sim \sqrt{2\pi m}(m/e)^m$:

$$\begin{aligned}\binom{n}{k} p^k (1-p)^{n-k} &\sim \frac{\sqrt{2\pi n}(n/e)^n}{\sqrt{2\pi k}(k/e)^k \sqrt{2\pi(n-k)}((n-k)/e)^{n-k}} p^k (1-p)^{n-k} \\ &= \frac{1}{\sqrt{2\pi}} \sqrt{\frac{n}{k(n-k)}} \left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k}.\end{aligned}$$

Let $k = np + \delta_n$, where $\delta_n = x_k\sqrt{np(1-p)}$. Since $|k - np| = O(\sqrt{n})$, $\delta_n/n \to 0$.
Notice that $\sqrt{\frac{n}{k(n-k)}} \sim \frac{1}{\sqrt{np(1-p)}}$. The term $\left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k}$ is then analyzed using Taylor expansions of logarithms, similar to:

$$\ln\left[\left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k}\right] = k\ln\left(1 - \frac{\delta_n}{np+\delta_n}\right) + (n-k)\ln\left(1 + \frac{\delta_n}{n(1-p)-\delta_n}\right).$$

Using the taylor series $\ln(1+u) \approx u - u^2/2$ and doing some manipulations, this logarithm simplifies to approximately $-\frac{\delta_n^2}{2np(1-p)} = -x_k^2/2$. Thus, $\left(\frac{np}{k}\right)^k \left(\frac{n(1-p)}{n-k}\right)^{n-k} \sim e^{-x_k^2/2}$. Combining these approximations yields the result. □

After this, proving the original result for this case isn't hard, and we are not including it here. The rough idea uses the fact that the de Moivre-Laplace Theorem approximates each $P(S_n = k)$ by a value proportional to the height of the standard normal curve at the corresponding standardized point $x_k = \frac{k-np}{\sigma_n}$. This sum then becomes a Riemann sum for the integral of the normal density function over the interval $[a, b]$ where $\frac{1}{\sigma_n}$ acts like the infinitesimal width dx.

Bibliography

[1] Ayan Bhattacharya. Lecture notes on probability ii. SI 537 IIT Bombay, Autumn 2024-25.

[2] E. Lesigne. *Heads or Tails: An Introduction to Limit Theorems in Probability*. Student mathematical library. American Mathematical Society, 2005.