LEC018 Review of Probability II

VG441 SS2021

Cong Shi
Industrial & Operations Engineering
University of Michigan

Limit Theorems

- WLLN

Let $X_1, \ldots, X_n$ be i.i.d. having $\mathbb{E}[X] = \mu$ and variance σ^2, then

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad \bar{X}_n - \mu \quad \xrightarrow{i.p.} \quad 0 \quad \text{as } n \to \infty.$$

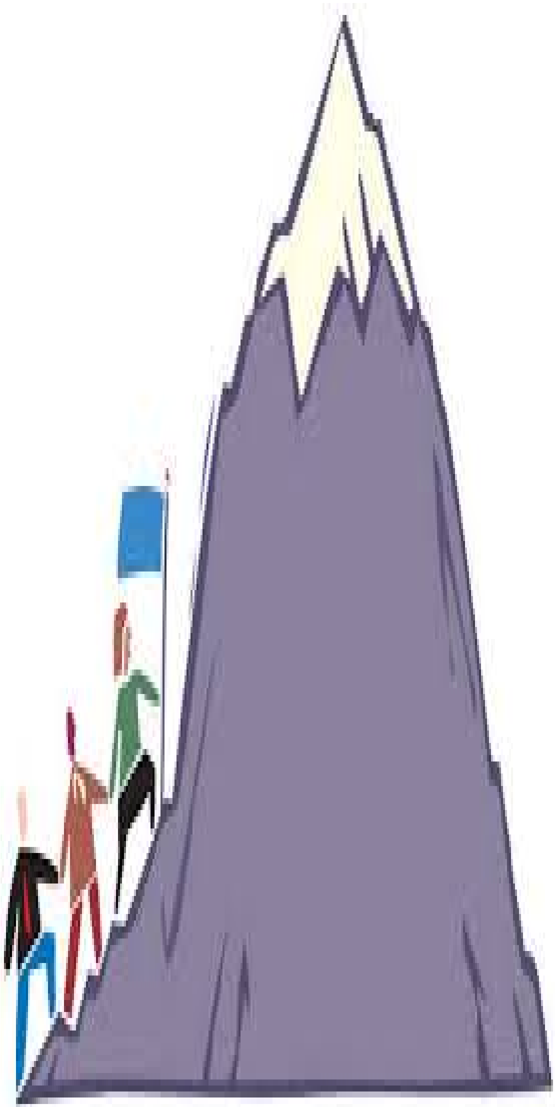
- CLT

Let $X_1, \ldots, X_n$ be i.i.d. having $\mathbb{E}[X_1] = \mu$ and variance σ^2, then

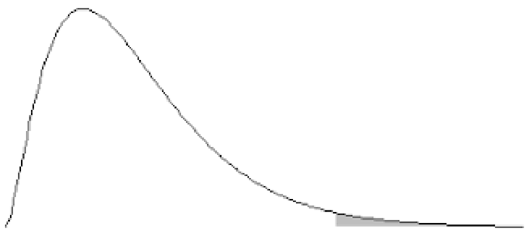
$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad \text{and} \qquad \sqrt{n}\left(\bar{X}_n - \mu\right) \quad \xrightarrow{d} \quad N\left(0, \sigma^2\right) \quad \text{as } n \to \infty.$$

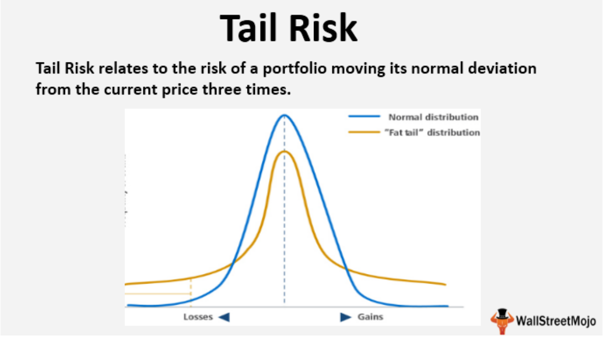
Limit Theorems

- Statistical snapshot at different levels



Tail Approximations

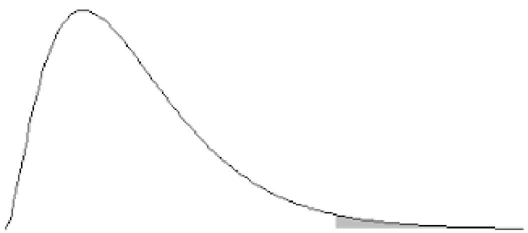




Tail Approximations

SLLN + CLT tells us $S_n \approx n\mu + \sqrt{n}\sigma N(0,1)$. So CLT handles deviation of size $\sqrt{n}$: $\mathbb{P}\left(S_n > n\mu + \delta\sqrt{n}\right) \approx \mathbb{P}(Z > \delta/\sigma)$ for moderate or large n.

A simple example:



Consider i.i.d. r.v.'s $X_1, \ldots, X_{16}$ where $X_i \sim U[0,1]$ for all $i = 1, \ldots, 16$. We want to bound

$$\mathbb{P}\left(\sum_{i=1}^{16} X_i \geq 10\right).$$

Let $S_{16} = \sum_{i=1}^{16} X_i$ and then $\mathbb{E}(S_{16}) = 16\mathbb{E}(X_1) = 8$. The true distribution is the Irwin-Hall distribution with n.

- Using Markov's inequality,

$$\mathbb{P}\left(S_{16} \geq 10\right) \leq \mathbb{E}(S_{16})/10 = 8/10 = 0.80.$$

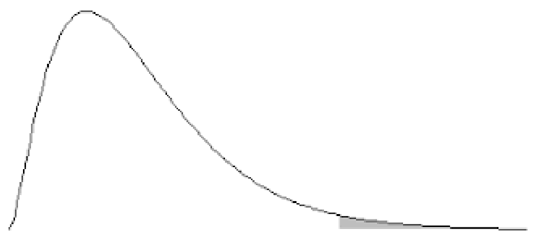
- Using Chebyshev's inequality,

$$\mathbb{P}\left(S_{16} \geq 10\right) = \frac{1}{2}\mathbb{P}\left(|S_{16} - 8| \geq 2\right) \leq \frac{1}{2}\left(\frac{1}{2^2}\right)\sigma^2_{S_{16}} = \frac{1}{8}\left(\frac{16}{12}\right) = 0.17.$$

- Using CLT, we have

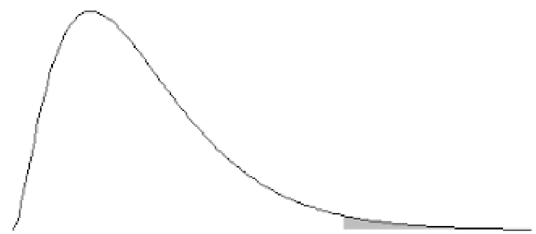
$$\mathbb{P}\left(S_{16} \geq 10\right) = \mathbb{P}\left(S_{16} \geq 8 + 0.5(4)\right) \approx \mathbb{P}\left(Z \geq \frac{0.5}{\sqrt{1/12}}\right) = 1 - \Phi(1.732) = 0.042.$$

Big Question



How many samples/experiments n do you need for the performance to be robust (having a tail that is small, preferably exponentially smaller as n increases)?

Big Question



How many samples/experiments n do you need for the performance to be robust (having a tail that is small, preferably exponentially smaller as n increases)?

CLT "roughly" handles deviation of size $\sqrt{n}$:

$$\mathbb{P}\left(S_n > n\mu + \delta\sqrt{n}\right) \approx \mathbb{P}(Z > \delta/\sigma) \quad \text{for moderate or large } n.$$

Large deviation "exactly" handles deviation of size n:

$$\mathbb{P}\left(S_n > n\mu + n\delta\right).$$

CLT is insufficient to bound the above in a sense that

$$\mathbb{P}\left(S_n > n\underbrace{(\mu+\delta)}_{a}\right) = \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} > \frac{\sqrt{n}(a-\mu)}{\sigma}\right), \quad \text{but} \quad \sqrt{n}\left(\frac{a-\mu}{\sigma}\right) \to \infty!$$

Concentration inequalities

Consider an i.i.d. sequence $X_1, X_2 \ldots$. Fix a value $a > \mu$ and fix a positive parameter $\theta > 0$. We have

$$\begin{aligned} \mathbb{P}\left(\frac{\sum_{1\le i\le n} X_i}{n} > a\right) &= \mathbb{P}\left(\sum_{1\le i\le n} X_i > na\right) = \mathbb{P}\left(e^{\theta \sum_{1\le i\le n} X_i} > e^{\theta na}\right) \\ &\le \frac{\mathbb{E}[e^{\theta \sum_{1\le i\le n} X_i}]}{e^{\theta na}} = \frac{\mathbb{E}[e^{\theta X_1} \cdots e^{\theta X_n}]}{(e^{\theta a})^n}. \end{aligned}$$

But recall that X_i's are i.i.d. Therefore $\mathbb{E}[e^{\theta X_1} \cdots e^{\theta X_n}] = (\mathbb{E}[e^{\theta X_1}])^n$. Thus, we obtain an upper bound

$$\mathbb{P}\left(\frac{\sum_{1\le i\le n} X_i}{n} > a\right) \le \left(\frac{\mathbb{E}[e^{\theta X_1}]}{e^{\theta a}}\right)^n.$$

First assume for a moment that $\mathbb{E}(\theta X_1)$ is finite for all θ in some interval $[0, \theta_0)$. Note that when $\theta = 0$ the ratio $\frac{\mathbb{E}[e^{\theta X_1}]}{e^{\theta a}} = 1$. Now differentiate this ratio with respect to θ at $\theta = 0$:

$$\frac{d}{d\theta}\frac{\mathbb{E}[e^{\theta X_1}]}{e^{\theta a}}\bigg|_{\theta=0} = \frac{\mathbb{E}[X_1 e^{\theta X_1}]e^{\theta a} - a e^{\theta a}\mathbb{E}[e^{\theta X_1}]}{e^{2\theta a}}\bigg|_{\theta=0} = \mathbb{E}[X_1] - a = \mu - a < 0.$$

Therefore, for sufficiently small θ the ratio ratio $\frac{\mathbb{E}[e^{\theta X_1}]}{e^{\theta a}}$ is less than unity!

Concentration inequalities

Given an i.i.d. sequence $X_1, \ldots, X_n$ suppose $\mathbb{E}[e^{\theta X_1}]$ is finite for all θ in some interval $[0, \theta_0)$. Let $a > \mu = \mathbb{E}[X_1]$. Then for some sufficiently small $\theta > 0$ there holds $\frac{\mathbb{E}[e^{\theta X_1}]}{e^{\theta a}} < 1$ and, moreover,

$$\mathbb{P}\left(\frac{\sum_{1 \le i \le n} X_i}{n} > a\right) \le \left(\frac{\mathbb{E}[e^{\theta X_1}]}{e^{\theta a}}\right)^n.$$

In other words, the large deviation probability is exponentially small.

One degree of freedom (that we can leverage)…

How small can we make this ratio? We have some freedom in choosing θ as long as $\mathbb{E}[e^{\theta X_1}]$ is finite. So we could try to find θ which minimizes the ratio $\frac{\mathbb{E}[e^{\theta X_1}]}{e^{\theta a}}$. This is what we will do. The surprising conclusion of the large deviations theory is that such a minimizing value θ^* exists and is tight. Namely it provides the *correct decay rate*!

Chernoff Bound

Arguably the most useful inequality in probability theory:

A Legendre transform of a r.v. X is the function $l(a) \triangleq \sup_\theta(\theta a - \log M(\theta))$.

We have established an upper bound on the probability of large deviations

$$\mathbb{P}\left(\frac{\sum_{1\le i\le n} X_i}{n} > a\right) \le e^{-l(a)n},$$

where $l(a)$ is the Legendre transform corresponding to the distribution of random variable X_1. This upper bound is **tight**!

Exponential Distribution Example

Exponential distribution with parameter λ. Recall that $M(\theta) = \lambda/(\lambda - \theta)$ when $\theta < \lambda$ and $M(\theta) = \infty$ otherwise. Therefore when $\theta < \lambda$,

$$l(a) = \sup_\theta \left(a\theta - \log \frac{\lambda}{\lambda - \theta} \right) = \sup_\theta \left(a\theta - \log \lambda + \log(\lambda - \theta) \right).$$

Setting the derivative of $g(\theta) = a\theta - \log \lambda + \log(\lambda - \theta)$ equal to zero we obtain the equation $a - 1/(\lambda - \theta) = 0$ which has the unique solution $\theta^* = \lambda - 1/a$. Therefore,

$$l(a) = a(\lambda - 1/a) - \log \lambda + \log(\lambda - \lambda + 1/a) = a\lambda - 1 - \log \lambda - \log a.$$

The large deviations bound then tells us that when $a > 1/\lambda$,

$$\mathbb{P}\left(\frac{\sum_{1 \le i \le n} X_i}{n} > a \right) \approx e^{-(a\lambda - 1 - \log \lambda - \log a)n}.$$

Say $\lambda = 1$ and $a = 1.2$. This approximation gives $\approx e^{-(0.2 - \log 1.2)n}$. Recall that the process $X_1, X_1 + X_2, \ldots, X_1 + \ldots + X_n, \ldots$ is a Poisson process with $\lambda = 1$. We can compute the probability $\mathbb{P}(\sum_{1 \le i \le n} X_i > 1.2n)$ exactly: it is the probability that the Poisson process has at most $n - 1$ events before time $1.2n$. Thus,

$$\mathbb{P}\left(\frac{\sum_{1 \le i \le n} X_i}{n} > 1.2 \right) = \mathbb{P}\left(\sum_{1 \le i \le n} X_i > 1.2n \right) = \sum_{0 \le k \le n-1} \frac{(1.2n)^k}{k!} e^{-1.2n}.$$

It is not at all clear how revealing this expression is. In hindsight, we know that it is approximately $e^{-(0.2 - \log 1.2)n}$.

Normal Distribution Example

Standard normal distribution. Recall that $M(\theta) = e^{\frac{\theta^2}{2}}$ when X_1 has the standard Normal distribution. The expected value $\mu = 0$. Thus we fix $a > 0$ and obtain

$$l(a) = \sup_{\theta}(a\theta - \theta^2/2) = a^2/2,$$

achieved at $\theta^* = a$. Again we see that $l(a)$ is (as it should be) a convex function of a. Thus for $a > 0$, the large deviations theory predicts that

$$\mathbb{P}\left(\frac{\sum_{1\le i\le n} X_i}{n} > a\right) \approx e^{-\frac{a^2}{2}n}.$$

Again we could compute this probability directly. We know that $\frac{\sum_{1\le i\le n} X_i}{n}$ is distributed as a Normal random variable with mean zero and variance $1/n$. Thus

$$\mathbb{P}\left(\frac{\sum_{1\le i\le n} X_i}{n} > a\right) = \frac{\sqrt{n}}{\sqrt{2\pi}} \int_a^\infty e^{-\frac{t^2 n}{2}} dt.$$

One could show that this integral is "dominated" by its part around a, namely, $\frac{\sqrt{n}}{\sqrt{2\pi}} e^{-\frac{a^2}{2}n}$. This is consistent with the large deviations theory. The lower order term $\frac{\sqrt{n}}{\sqrt{2\pi}}$ disappears in the approximation on the log scale.

Poisson Distribution Example

Poisson distribution. Suppose X has a Poisson distribution with parameter λ. Recall that $M(\theta) = e^{e^{\theta\lambda} - \lambda}$. Then

$$l(a) = \sup_{\theta}(a\theta - (e^{\theta}\lambda - \lambda)).$$

Setting derivative to zero we obtain $\theta^* = \log(a/\lambda)$ and $l(a) = a\log(a/\lambda) - (a - \lambda)$.

In this case as well we can compute the large deviations probability explicitly. The sum $X_1 + \ldots + X_n$ of Poisson random variables is also a Poisson random variable with parameter λn. Therefore

$$\mathbb{P}\left(\sum_{1 \le i \le n} X_i > an\right) = \sum_{m > an} \frac{(\lambda n)^m}{m!} e^{-\lambda n}.$$

But again it is hard to infer a more explicit rate of decay using this expression.

Hoeffding Inequality

Let $X_1, \ldots, X_n$ be i.i.d. random variables on a bounded support $[a, b]$. Let $\mathbb{E}(X_1) = \mu$. Then, for any $\epsilon > 0$,

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

Example:

If $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$, then

$$\mathbb{P}(|\bar{X}_n - p| > \epsilon) \leq 2e^{-2n\epsilon^2}.$$

Hoeffding Inequality

Let $X_1, \ldots, X_n$ be i.i.d. random variables on a bounded support $[a, b]$. Let $\mathbb{E}(X_1) = \mu$. Then, for any $\epsilon > 0$,

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

Proof. WLOG assume $\mu = 0$.

$$\mathbb{P}(X_1 + \ldots + X_n \geq n\epsilon) = \mathbb{P}(e^{t(X_1+\ldots+X_n)} \geq e^{tn\epsilon}) \leq \frac{\mathbb{E}[e^{t(X_1+\ldots+X_n)}]}{e^{tn\epsilon}} = \frac{(\mathbb{E}e^{tX_1})^n}{e^{tn\epsilon}}.$$

We bound the MGF of X_1. Below is a very useful inequality: for any X with zero mean and bounded support $[a, b]$,

$$\mathbb{E}e^{tX} \leq e^{t^2(b-a)^2/8}. \tag{1}$$

We will show this later. Now we have

$$\mathbb{P}(X_1 + \ldots + X_n \geq n\epsilon) \quad \leq \quad \inf_t \frac{(\mathbb{E}e^{tX_1})^n}{e^{tn\epsilon}} \leq \inf_t \left(e^{nt^2(b-a)^2/8 - tn\epsilon}\right).$$

Choose $t = 4\epsilon/(b-a)^2$ such that the exponent is minmized. This completes the proof. $\square$

Bound the MGF

We then prove that our claim is true, i.e., for any X with zero mean and bounded support $[a, b]$,

$$\mathbb{E}e^{tX} \le e^{t^2(b-a)^2/8}. \tag{1}$$

We first write $X = \frac{b-X}{b-a}a + \frac{X-a}{b-a}b$. By convexity,

$$\begin{aligned} \mathbb{E}e^{tX} &\le \mathbb{E}\left(\frac{b-X}{b-a}e^{ta}\right) + \mathbb{E}\left(\frac{X-a}{b-a}e^{tb}\right) = \frac{b}{b-a}\left(e^{ta}\right) - \frac{a}{b-a}\left(e^{tb}\right) \\ &= e^{at} + e^{at}\left(\frac{a}{b-a}\right) - e^{bt}\left(\frac{a}{b-a}\right) = e^{at}\left(1 + \left(\frac{a}{b-a}\right) - \left(\frac{a}{b-a}\right)e^{t(b-a)}\right) = e^{g(u)}, \end{aligned}$$

where

$$g(u) = -\gamma u + \log(1 - \gamma + \gamma e^u), \quad \gamma = -\frac{a}{b-a}, \quad u = t(b-a).$$

Note that $g(0) = g'(0) = 0$ and $g''(x) \le 1/4$ for all $x > 0$. Using Taylor's Theorem, we have for some $\xi \in [0, u]$,

$$g(u) = g(0) + ug'(0) + \frac{u^2}{2}g''(\xi) \le u^2/8 = t^2(b-a)^2/8.$$

Hoeffding Inequality

Let $X_1, \ldots, X_n$ be i.i.d. random variables on a bounded support $[a, b]$. Let $\mathbb{E}(X_1) = \mu$. Then, for any $\epsilon > 0$,

$$\mathbb{P}(|\bar{X}_n - \mu| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$

Corollary:

Let $X_1, \ldots, X_n$ be i.i.d. random variables on a bounded support $[a, b]$. Let $\mathbb{E}(X_1) = \mu$. Then

$$|\bar{X}_n - \mu| \leq \sqrt{\frac{(b-a)^2}{2n} \log\left(\frac{2}{\delta}\right)}, \qquad \text{with probability at least } 1 - \delta.$$