9.2 LAW OF LARGE NUMBERS

For $S_n \sim \mathrm{Bin}(n,p)$

we used the CLT to figure out that

$$\lim_{n\to\infty} \mathbb{P}\left(\left| \frac{S_n}{n} - p \right| \leq \epsilon \right) = 1.$$

The reason this was true was because

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

independent Bernoullis.

Since they are all Bernoullis, we usually say that $X_1, \dots, X_n$ are IID random variables.

IID : Independent, identically distributed.

Thm: Let $X_1, \dots X_n$ be iid rvs with mean μ and variance σ^2. Let

$$S_n = \frac{X_1 + \cdots X_n}{n}$$

Then

$$\lim_{n\to\infty} \mathbb{P}\left(\left| \frac{S_n}{n} - \mu \right| \geqslant \epsilon \right) = 0.$$

lets check first that

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

$$\begin{aligned} \mathrm{Var}(S_n) &= \mathrm{Var}(X_1 + \cdots + X_n) \\ &= \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n) \\ &= n\sigma^2 \end{aligned}$$

Pf:

$$\mathbb{P}\left(\left| \frac{S_n}{n} - \mu \right| \geqslant \epsilon \right) = \mathbb{P}\left(|S_n - n\mu| \geqslant n\epsilon \right)$$

$$\leq \frac{\mathrm{Var}(S_n)}{n^2 \epsilon^2} = \frac{\cancel{n}\sigma^2}{n^{\cancel{2}}\epsilon^2} \to 0.$$

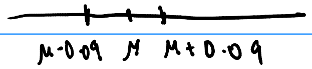
↑

Chebyshev inequality. If we didnt learn the inequality we may skip the proof.

$\epsilon = 0.1$

$G = P\left(\left| \frac{S_n}{n} - \mu \right| < 0.1 \right)$ $\qquad \epsilon = 0.09$

$H = P\left(\left| \frac{S_n}{n} - \mu \right| < 0.09 \right)$



POLL

Which prob is greater.

A $\qquad\qquad$ B

$G \le H$ $\qquad\qquad$ $H \le B$

$$\underset{\overset{\shortparallel}{1}}{\lim_{n \to \infty} P\left(\left| \frac{S_n}{n} - \mu \right| < 0.09 \right)} \le \underset{\overset{\shortparallel}{1}}{\lim_{n \to \infty} P\left(\left| \frac{S_n}{n} - \mu \right| < 0.1 \right)}$$

You have to choose a larger n to get a higher prob. of being closer to the mean.

Ex: Suppose we work in a factory that produces batteries. Workers flag products as defective if they dont measure the correct voltage as they come off the line. If more than 10 defective batteries come off the line in a day, production is stopped for the day, and you figure out the problem.
(muri, mura, muda)
↳ Toyota Way

We want to estimate the mean # of defective items produced in a day to with 0.1 and know this with high probability.
↳ $\epsilon =$

→ # of failures after n days

Let $S_n = X_1 + \cdots + X_n$

↳ # of failures on 1st day.

$$P\left(\left|\frac{S_n}{n} - \mu\right| > 0.1\right) \leq \frac{Var(S_n/n)}{(0.1)^2} = \frac{1}{n^2}\frac{Var(S_n)}{(0.1)^2}$$

$$= \frac{1}{n^{\cancel{2}}}\frac{\cancel{n}\,Var(X_1)}{(0.1)^2}$$

How do we estimate $Var(X_1)$?

$X \leq 10$ $\qquad X^2 \leq 100$ $\qquad \Rightarrow \quad Var(X) \leq$

$\Rightarrow$

$$P\left(\left|\frac{S_n}{n} - \mu\right| > 0.1\right) \leq$$

Suppose we wanted to be ______ %
sure that we're within 0.1 of the mean # of failed batteries.

Then how many days would we need to observe the production?

Central limit theorem:

Lets take an example: $\{X_i\}_{i=1}^{50}$ are iid rvs representing, say student scores.

Then $0 \leq X_i \leq 100$.

Suppose the average student has score

$$\mathbb{E}[X_i] = 60$$

and the standard deviation $\sigma = 13$

Then, if

$$S_{50} = X_1 + \cdots + X_{50}$$

then $\mathbb{E}[S_n] = 50\,\mathbb{E}[X_1] = 50 \cdot 60$

In fact $\frac{S_n}{n} \to 60.$ as n gets large.

So this means that $S_{50} \approx 60 \cdot 50$,

but it wont be exactly equal to 3000.

$$S_{50} = X_1 + X_2 + \cdots + X_{50}$$

assume student scores are independent and identically dist.

$$E[S_{50}] = E[X_1 + X_2 + \cdots + X_{50}]$$

$$=$$

$$= 50\,E[X_1] = 50 \cdot 60$$

↳ avg. score.

What does the LAW of LARGE NUMBERS TELL US?

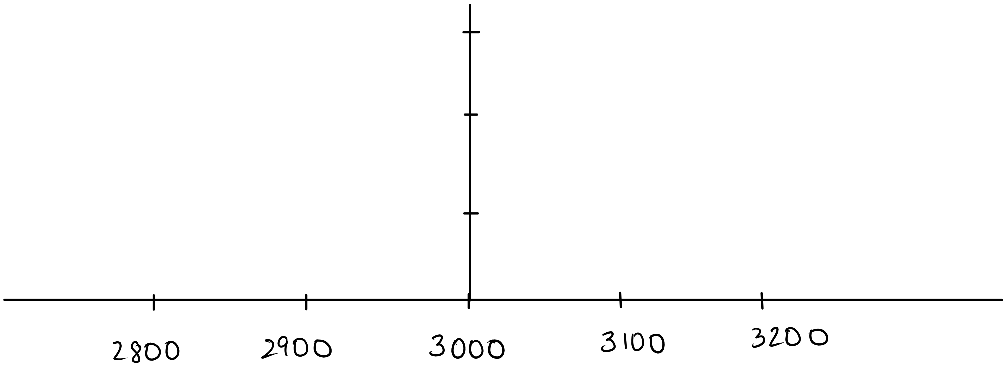
$$P\left(\left|\frac{S_{50}}{50} - E[X_1]\right| > \epsilon\right) \approx$$

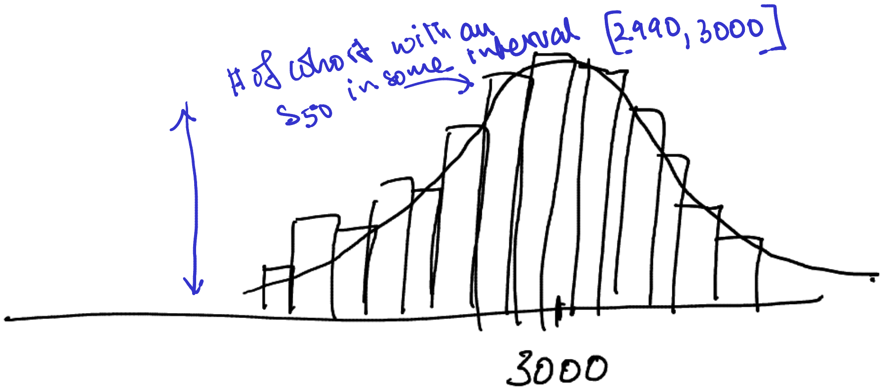
We could also estimate this using Chebyshev:

But what if you wanted to estimate:

$$P(a \leq S_{50} \leq b)$$

Suppose you teach this class 5 times: Then you will get scores like : 2901, 3120, 3250, 2890 for S_{50}.





If you have different sets of 50 students, you will get a bunch of different values for S_n:

2990, 2870, 3100, 3150, 4000,

You can use there values to form a histogram.

We will get a histogram that is awfully similar to a normal distribution.

What is the "width" of the (or limiting curve (more precisely)) histogram?

Controlled by $\sqrt{Var(S_n)}$ = standard deviation of S_{50}

$Var(S_n) = n \, Var(X_1) \Rightarrow \sqrt{Var \, S_n} = \sqrt{n\sigma^2}$

whatever σ^2 is for the class.

$$\begin{aligned} Var(S_{50}) &= Var(X_1 + X_2 + \cdots + X_{50}) \\ &= Var(X_1) + \cdots + Var(X_{50}) = 50\, Var(X_1) = 50\sigma^2 \\ &= 50 \cdot (13)^2 \end{aligned}$$

Central Limit Theorem says.

$$P\left(a \leq \frac{S_{50} - E[S_{50}]}{\sqrt{Var(S_{50})}} \leq b \right) \approx \Phi(b) - \Phi(a)$$

Theorem: Suppose $\{X_i\}_{i=1}^{n}$ are iid random variables with mean μ and variance σ^2.

Then

$$\lim_{n\to\infty} P\left(\frac{S_n - \mu n}{\sqrt{n\sigma^2}} \leq t \right) = \Phi(t)$$

Previously we had seen the BINOMIAL CLT. There,

$\{X_i\}_{i=1}^{n}$ were iid Bernoulli (p)

rvs.

$E[X_1] = p \qquad Var(X_1) = p(1-p)$

So we had

$$P\left(\frac{S_n - np}{\sqrt{np(1-p)}} \le t \right) \longrightarrow \Phi(t)$$

The CLT we have stated now holds for a large class of rvs.

- De Moivre- Laplace just for Binomials.

- Late 1800s & early 1900s Lindeberg, Lyapunov to all iid sums of rvs with finite mean & variance.

Ex: We roll 1000 dice and add up the values on their faces. What is the probability that the sum is larger than 3600?

X_i = value on i^{th} die

$S_{1000} = X_1 + X_2 + \cdots + X_{1000}$

$P(\qquad\qquad)$

$E[S_{1000}] =$

$Var(S_{1000}) =$

$$P\left(S_{1000} \geq 3600\right) = P\left(S_{1000} - E[S_{1000}] \geq 3600 - 3500\right)$$

$$= P\left(\frac{S_{1000} - E[S_{1000}]}{\sqrt{Var(Sn)}} \geq \frac{3600 - 3500}{\sqrt{2920}}\right)$$

$$= P\left(\frac{S_{1000} - E[S_{1000}]}{\sqrt{Var(Sn)}} \geq \frac{100}{\sqrt{2920}}\right)$$

$$\approx 1 - \Phi\left(\frac{100}{\sqrt{2920}}\right)$$

Ex:

Apple opens a new store and offers an ipad for its 1000th customer.
Whats the probability that the 1000th customer will arrive within a 7 days given that arrival times between customers are Exp(λ), $\lambda = 100 \ {}^{1}/_{days}$

T_i = arrival time of n^{th} customer

$$\{ \qquad \qquad \qquad \} = \{1000 \text{ customers within } 7 \text{ days}\}$$

Let $S_n =$

$$P(\qquad \qquad \qquad)$$

Center and scale appropriately: