The normal distribution can sometimes approximate probabilities for linear combinations of random variables

Central Limit Theorem

If X1, X2, …, Xn are all independent random variables *with the same distribution*

- Mean μ
- Variance σ^2

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}}$$

Approaches N(0,1) as n -> ∞

CLT works for all distributions except those whose mean and variance *do not exist* (not finite)

Approximation is better as n gets bigger

If the distribution is symmetric, the approximation is also probably better

n ≥ 30 is a good general rule of thumb for the number of samples needed to approximate a normal distribution

- Not set in stone — this just comes up a lot
- "Approximate" — distribution is not necessarily exactly normal

If X1, X2, …, Xn are normally distributed, then S_n and X̄ have *exact normal distributions* for any value of n (since we're just adding variables that are *already* normal)

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \sigma^2/n)$$

However, if they are not normally distributed, then S_n and X̄ will *approximate* a normal distribution

Ex. Suppose fires reported to a fire station satisfy the conditions for a Poisson process, with a mean of 1 fire every 4 hours.

Find the approximate probability that the 500th fire of the year is reported on the 84th day of the year.

Let X_i = the time between the (i-1)th fire and the i-th fire

X1 is the time to the first fire

We have to wait 4 hours per fire

So θ = 4 hours = 1/6 day

X_i ~ Exponential(θ = 1/6)

$$S_{500} = \sum_{i=1}^{500} X_i$$: time (in days) until 500th fire

We want $P(83 < S_{500} \leq 84)$

By the CLT, $S_n \sim N(n\mu, n\sigma^2)$

$\mu = \theta$; $\sigma^2 = Var(S_n) = \theta^2$

$\Rightarrow S_{500} \sim N\left(\frac{500}{6}, \frac{500}{36}\right)$

Translating to z-scores, we have

$P(-0.09 < z \leq 0.18)$
$= 0.10728$

Actual answer is 0.1063945 — approximation is close because n=500 is sufficiently large

Note: when using the CLT (a normal distribution) to approximate a discrete distribution, we must adjust our answer slightly by doing *continuity correction*

Ex. Suppose $X \sim \text{Bin}(100, 0.5)$

We want to use the CLT to approximate $P(X=50)$

If we didn't apply continuity correction:

- $X=50$
- $\mu=50$
- These would cancel out when translating to normal, and $P(X=50)$ would be 0, which is wrong

Instead, we can use:

$P(X=50) = P(49 < X < 51)$

Then meet in the middle (compromise between the two above): $P(49.5 < X < 50.5)$

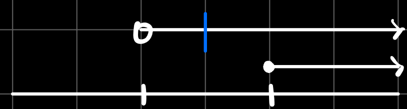
correction

Ex. $P(X < 15)$

- This is $P(X \le 14)$
- Then compromise between the two: $P(X \le 14.5)$

Ex. $P(X \le 12)$?

- Equal to $P(X < 13)$ -> $P(X \le 12.5)$



Ex. $P(X \ge 6)$

- Equal to $P(X > 5)$ -> $P(X \ge 5.5)$

Using the normal distribution to approximate the binomial distribution:

$$z = \frac{X - np}{\sqrt{np(1-p)}}$$

Ex. $X \sim \text{Bin}(20, 0.4)$. Use the normal distribution to approximate $P(4 \le X \le 12)$

This is equal to $P(3.5 \le X \le 12.5)$

$= P(X \le 12.5) - P(X \le 3.5)$

$$= P\left(Z \le \frac{\overset{np}{12.5 - 8}}{\sqrt{(20)(0.4)(0.6)}}\right) - \ldots$$

$= 0.95964$

Using binomial gives us 0.96301 → very close