Central Limit Theorem

This is from Ch 7 of our free coursenotes but we are doing this ahead of Ch 6 because it has something to do with normal distributions (Ch 5), and we are in a good position now to learn this very important theorem.

① It is important for students to remember that when we take a random sample of size n from a distribution X, we write the observations as

$$X_1, X_2, X_3, ..., X_n$$

② Let's say $X_1, X_2, ..., X_n$ is a random sample of size n from any distribution with mean μ and variance σ^2.

③ Say sample size n is large (≥ 20). Then

$$\boxed{\bar{X} \overset{approx.}{\sim} N(\mu, \sigma^2/n)}$$

Note: In some problems, the question is about sum , not average. You can still use CLT. How?

Let T be sum of $X_1, X_2, ..., X_n$

$$T = \sum_{i=1}^{n} X_i \quad and \quad \bar{X} = \frac{\sum X_i}{n}$$

$$\Rightarrow T = n\bar{X}$$

$$E(T) = E(n\bar{X})$$

$$= n\mu$$

$$\begin{aligned} Var(T) &= Var(n\bar{X}) \\ &= n^2 Var(\bar{X}) \\ &= n^2 \frac{\sigma^2}{n} \\ &= n\sigma^2 \end{aligned}$$

sum

$$\boxed{T \overset{approx.}{\sim} N(n\mu, n\sigma^2)}$$

How large should n be to apply CLT?

need larger n if skewed

It depends on how asymmetrical X is.

Your course notes say in many practical situations, CLT normal approximation can be used when $n \geq 20$. (some books say $n \geq 30$).

Let's work on Ex.7.1 (Pg 120) in course notes.

Example 7.1 A system consist of 25 independent parts connected in such a way that the i^{th} part automatically turns-on when the $(i-1)^{th}$ part burns out. The expected lifetime of each part is 10 weeks and the standard deviation is equal to 4 weeks. (a) Calculate the expected lifetime and standard deviation for the system. (b) Calculate the probability that the system will last more than its expected life. (c) Calculate the probability that the system will last more than 1.1 times its expected life. (d) What are the (approximate) median life and interquartile range for the system?

(a) Let X_i denote the lifetime of the i^{th} component

$$X_i \sim \quad ? \qquad \textit{Distribution unknown}$$

$$E(X_i) = 10$$
$$Var(X) = 4^2$$

(Idea: Suspect CLT needed somewhere in problem since distribution unknown)

Let T be lifetime of entire system

$$T = X_1 + X_2 + ... + X_{25}$$

$$\begin{aligned}
E(T) &= E(X_1 + X_2 + ... + X_{25}) = 25E(X_1) \\
&= 25 \times 10 = 250 \\
Var(T) &= Var(X_1 + X_2 + ... + X_{25}) \\
&= 25Var(X_1) \quad [X_i's \textit{ are independent}] \\
&= 25 \times 16 = 400 \\
SD(T) &= 20
\end{aligned}$$

(b) Note:

CLT says $\bar{X} \sim N(\overset{\mu}{10}, \overset{\frac{\sigma^2}{n}}{\frac{16}{25}})$

Alternatively, $T \sim N(250, 25 \times 16)$

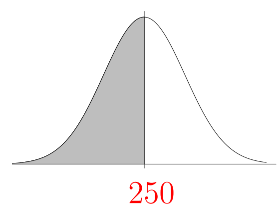
$$T \sim N(\underset{n\mu}{250}, \underset{n\sigma^2}{400})$$

For part(b), we use T since the question wants $P(T > E(T)) = P(T > 250) = \frac{1}{2}$ [think: do we need normal table in this case?]

(c)

$$\begin{aligned} P(T > 1.1E(T)) &= P(T > 275) \\ &= P(Z > \frac{275 - 250}{\sqrt{400}}) \\ &= P(Z > 1.25) \\ &= 1 - P(Z < 1.25) \\ &= 1 - 0.8944 \\ &= 0.1056 \end{aligned}$$

(d) Since $T \sim N(250, 400)$, Median=250 [think:



]

Let q_1 be lower quartile

$$\begin{aligned} P(T < q_1) &= 0.25 \\ P(Z < \frac{q_1 - 250}{20}) &= P(Z < -0.675) \\ \frac{q_1 - 250}{20} &= -0.675 \\ q_1 &= 20(-0.675) + 250 \\ &= 236.5 \end{aligned}$$

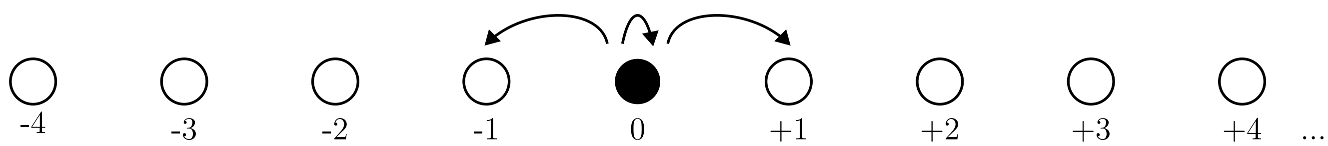
Let q_3 be upper quartile

$$\begin{aligned}
P(T < q_3) &= 0.75 \\
P(Z < \frac{q_3 - 250}{20}) &= P(Z < 0.675) \\
\frac{q_3 - 250}{20} &= 0.675 \\
q_3 &= 250 + (0.675 \times 20) \\
&= 263.50 \\
IQR &= q_3 - q_1 \\
&= 263.5 - 236.5 \\
&= 27
\end{aligned}$$

Example Random Walk Harder example on application of CLT

Physicists use random walks to model the process of diffusion, or random motion of particles. The position S_n of a particle at time n can be thought of as a sum of displacements $X_1, X_2, ...X_n$. Assuming the displacements are independent and identically distributed, we can use CLT to solve some questions on random walk.

Suppose at each step, a particle moving on sites labelled by integer is equally likely to move one step to the right, or one step to the left, or stay where it is.



Find approximately the probability that after 10,000 steps, the particle ends up

(a) more than 50 sites to the right of starting point?

(b) more than 50 sites to the left of starting point?

(c) more than 100 sites to the right of starting point?

<u>Sol:</u> Let X represent a single step, taking on a value of 1 if it moves to the right, -1 if it moves to the left or 0 if it stays put.

X	-1	0	1
P(X=x)	1/3	1/3	1/3

Note that X is discrete.

$$E(X) = 0 \qquad \left(since\ (-1 \times \frac{1}{3}) + (0 \times \frac{1}{3}) + (1 \times \frac{1}{3}) = 0\right)$$

$$\begin{aligned} Var(X) &= E(X^2) - [E(X)]^2 \\ &= \left[(-1)^2\frac{1}{3} + 0^2\frac{1}{3} + 1^2\frac{1}{3}\right] - 0^2 \\ &= (\frac{1}{3} + 0 + \frac{1}{3}) - 0^2 \\ &= \frac{2}{3} \end{aligned}$$

The problem in (a) is to find $P(S_{10,000} > 50)$

But

$$\begin{aligned} S_{10,000} &= X_1 + X_2 + X_3 + ... + X_{10,000} \quad (where\ X_i\ is\ i^{th}\ step) \\ E(S_{10,000}) &= E(X_1 + X_2 + X_3 + ... + X_{10,000}) \\ &= 10000E(X) \\ &= 0 \\ Var(S_{10,000}) &= Var(X_1 + X_2 + ... + X_{10,000}) \\ &= 10000Var(X) \\ &= 10000 \times \frac{2}{3} \\ &= \frac{20000}{3} \end{aligned}$$

It is wrong to think of $S_{10,000} = 10,000X$

By CLT (note n=100 is large)

$$S_{10000} \overset{approx.}{\sim} N\big(0, \frac{20000}{3}\big)$$

(a)

$$\begin{aligned} P(S_{10,000} > 50) &= P\big(Z > \frac{50-0}{\sqrt{20000/3}}\big) \\ &= P(Z > 0.61) \\ &= 1 - P(Z \leq 0.61) \\ &= 1 - 0.7291 \\ &= 0.2709 \end{aligned}$$

(b)

$$\begin{aligned} P(S_{10,000} < -50) &= P\big(Z < \frac{-50-0}{\sqrt{20000/3}}\big) \\ &= P(Z \leq \frac{-50}{\sqrt{20000/3}}) \\ &= P(Z \leq -0.61) \\ &= 0.2709 \end{aligned}$$

(c)

$$\begin{aligned} P(S_{10,000} > 100) &= P\big(Z > \frac{100-0}{\sqrt{20000/3}}\big) \\ &= P(Z > 1.22) \\ &= 1 - P(Z \leq 1.22) \\ &= 1 - 0.8888 \\ &= 0.1112 \end{aligned}$$

Alternative solution

You may convert the question to one on average.

(a) $P(S_{10,000} > 50) = P(\bar{X} > \frac{50}{10000})$

think $\frac{\sigma^2}{n}$

think of dividing both sides by 10000

$$\bar{X} \sim N(0, \frac{2/3}{10000})$$

$$\bar{X} \sim N(0, \frac{2}{30000})$$

$$\begin{aligned} P(\bar{X} > 0.005) &= P(Z > \frac{0.005 - 0}{\sqrt{2/30000}}) \\ &= P(Z > 0.61) \\ &= \textit{same answer as previous method} \end{aligned}$$

(b) & (c) can be done using averages as well.