Central Limit Theorem – the Meaning and the Usage

Convention about notation.
We are using notation “$X \sim N(\mu, \sigma^2)$” in lieu of saying that $X$ is a normal random variable with mean $\mu$ and standard deviation $\sigma$.

***

Assume a sample of $n$ measurements (or counts) that are distributed with the same probability distribution that has mean $\mu$ and standard deviation $\sigma$.

Example:
Take a sample of 30 GPAs of high school seniors.
Now look at the average $\bar{x}$ of your sample. It has a particular value.

| 3.23 | 3.11 | 2.35 | 1.39 | 3.80 | 3.12 | 2.35 | 2.76 | 3.43 | 1.91 | 2.72 | 2.23 | 2.10 | 3.24 | 3.55 |
| 2.34 | 3.78 | 1.66 | 3.56 | 3.45 | 2.87 | 3.10 | 3.07 | 3.28 | 1.64 | 3.01 | 3.45 | 3.10 | 2.24 | 1.98 |

Average (sample mean) $\bar{x} = 2.79$.

Take another sample of equal size and average of that one has another value.

| 3.29 | 2.21 | 2.78 | 2.19 | 3.50 | 3.30 | 2.35 | 2.75 | 3.23 | 2.82 | 2.22 | 2.23 | 2.10 | 3.24 | 3.45 |
| 2.34 | 2.35 | 2.45 | 2.56 | 3.45 | 2.87 | 3.40 | 3.43 | 2.28 | 1.89 | 3.11 | 2.55 | 3.32 | 2.55 | 1.67 |

Average $\bar{x} = 2.73$.

And so on.

The sample average $\bar{x}$ itself is a random variable with a probability distribution. We shall honor that fact by writing $\overline{X}$.

What kind of probability distribution?

**Fact A.**
If the original measurement $X$ is (approximately) normally distributed then the sample mean $\overline{X}$ is (approximately) normally distributed with the mean $\mu$ and the standard deviation $\frac{\sigma}{\sqrt{n}}$ (i.e., variance $\frac{\sigma^2}{n}$), if the sample size is $n$. 
But even if the original measurement is not normally distributed we still have:

**Fact B.**  
*Central Limit Theorem:*

As the sample size \(n\) increases the closer to the normal random variable the sample mean \(X\) is, regardless of the nature of the initial random variable. This normal is with mean \(\mu\) and the standard deviation \(\frac{\sigma}{\sqrt{n}}\) (i.e., variance \(\frac{\sigma^2}{n}\)).

The rule of thumb we use for the safe application of CLT is when the sample size is at least 30 or more.

This allows us to use normal distribution to assess probabilities of sample mean being in some particular range.

**Note:** The number \(\frac{\sigma}{\sqrt{n}}\) is often called *standard error of the mean.*

**Example 1.**

The height \((X)\) of 18-year-old men is approximately normally distributed with the mean 70 inches and standard deviation 3 inches, i.e., \(X \sim N(70, 3^2)\).

(a) What is the chance that an 18-year old man selected at random is between 67 and 73 inches tall?

Answer: \(P(67 \leq X \leq 73) \approx 0.683\) by using empirical rule.

(b) If we select the sample of nine 18-year-olds what is the chance that the mean height \(\bar{X}\) of the sample is between 67 and 73 inches?

Answer: Using the Fact A above we have that \(\bar{X}\) is \(N\left(70, \frac{3^2}{9}\right) = N(70, 1)\), therefore \(P(67 \leq \bar{X} \leq 73) \approx 0.997\) by empirical rule again.

**Example 2.**

The white blood cell count per cubic millimeter of blood is approximately normal with mean \(\mu = 7500\) and standard deviation \(\sigma = 1750\).

Doctor uses two-test average \(\bar{X}\). What is the chance that \(\bar{X} < 4500\)?
Answer:

Using Fact A above we have that $\bar{X}$ is approximately normal $N\left(7500, \frac{1750^2}{2}\right)$ thus

$$P\left(\bar{X} < 4500\right) \approx P\left(Z < \frac{4500 - 7500}{\frac{1750}{\sqrt{2}}}\right) \approx P\left(Z < -2.42\right) = 0.0078.\)$$

This is very unlikely, in less than 1% of the tests in general population.

Example 3.
Average Math SAT in the US for college bound seniors in 2004 was 518 with standard deviation of 118. We assume approximately normal distribution of the scores.
Professor Witt advises 4 freshmen. Let $\bar{X}$ be their average Math SAT.

What is the chance that this is

(a) better than 600?

(b) less than 400?

(c) better than 500 but less than 600?

Answers: First note that $\bar{X}$ is approximately normal, $N\left(518, \frac{118^2}{4}\right) = N\left(518, 59^2\right)$.

(a)

$$P\left(\bar{X} > 600\right) \approx P\left(Z > \frac{600 - 518}{\frac{118}{\sqrt{4}}}\right) \approx P\left(Z > 600 - 518\right) \approx P\left(Z > 1.39\right) = 0.0823.\)$$

(b)

$$P\left(\bar{X} < 400\right) \approx P\left(Z < \frac{400 - 518}{59}\right) \approx P\left(Z < -2\right) = 0.0228.\)$$

(c)

$$P\left(500 < \bar{X} < 600\right) \approx P\left(\frac{500 - 518}{59} < Z < \frac{600 - 518}{59}\right) \approx P\left(-0.31 < Z < 1.39\right) = 0.5394.\)$$
Example 4.

Suppose the population of adult males had a mean weight of 182 pounds with a standard deviation of 23 pounds. The number of males treated for heart related problems in local healthcare facility was 53 in 2011. Their average weight was 191. Would this lead to conclusion that males with heart related problems have more weight?

Answer:

By the Fact B from CLT we have that \( \bar{X} \) is approximately normal \( N \left( 182, \frac{23^2}{53} \right) \).

If we assume that the mean weight of 182 applies to this group of individuals then the chance that the weight would be above 191 is

\[
P(\bar{X} > 191) \approx P \left( Z > \frac{191 - 182}{\frac{23}{\sqrt{53}}} \right) \approx P(Z > 2.8487) \approx 0.002.
\]

This indicates that it is very unlikely that mean weight of 182 and standard deviation of 53 is applicable on this group. It is far more likely that the men with heart related problems are heavier than 182 pounds on average.

Example 5.

The mean lifetime of a certain type of incandescent light bulb is 900 hours with a standard deviation of 900 hours. We have a batch of 100 of these light bulbs. What is the chance these light bulbs are going last more than 100,000 hours?

Answer:

Let’s denote the lifetime of a single light bulb with \( X \). The question is posed as

\[
P \left[ X_1 + X_2 + \ldots + X_{100} > 100000 \right]
\]

where \( X_i \) means the lifetime of \( i \)-th light bulb. We can write this question as

\[
P \left[ \frac{X_1 + X_2 + \ldots + X_{100}}{100} > 1000 \right] = P\left[ \bar{X} > 1000 \right].
\]

Now we can use CLT because \( \bar{X} \) is approximately \( N \left( 900, \frac{900^2}{100} \right) \). That means normal with mean 900 and standard deviation 90.
The batch of bulbs has about 13% chance to last more than 100,000 hours.

Example 6.

A shipping company charges extra if the average shipping load during one week period exceeds the company limit of 20 tons per container.
Meat packing company ships on average 18.5 tons per container with a standard deviation of 7.4 tons. If they shipped 68 containers this week, what is the chance the average load was above 20 tons?

Answer:
The shipping load per container is a random variable \( X \) we do not know the nature of, apart from mean and standard deviation. Therefore to answer the question, what is the probability that \( \bar{X} \), the average of 68 shipping loads exceeds 20 tons we use CLT.
Therefore \( \bar{X} \) is approximately normal, \( N\left(18.5, \frac{7.4^2}{68}\right) \).

\[
P\left(\bar{X} > 20\right) \approx P\left(Z > \frac{20 - 18.5}{\frac{7.4}{\sqrt{68}}}\right) \approx P\left(Z > 1.67\right) = 0.0475
\]

The answer is approximately 5%.

Homework: Check online.
**Binomial Distribution by Normal Approximation**

(Optional)

Reminder:

\[ X = B(n, p) \] the count of successes for binomial distribution with the number of independent identical trials \( n \) and probability of success on one trial \( p \). Every trial can end with only two possible outcome – success and failure. Probability of failure is \( q = 1 - p \).

The probability mass formula is

\[
P[X = r] = \binom{n}{r} p^r q^{n-r}
\]

According to Central Limit Theorem, binomial distribution in case of a large \( n \) can be approximated with a normal distribution with the same mean \((np)\) and the same standard deviation \((\sqrt{npq})\). The implication is that respective probabilities for a binomial random variable (count of successes) can be approximated by using \( z \) – values of the variable in the charts for standard normal distribution.

There are two requirements the approximation needs to meet: \( np > 5, \ nq > 5 \).

Example 1.

1 out of 4 women that smoke will die from smoking related diseases. What is the probability that out of 1,000 women smokers more than 250 will die from smoking related diseases? What is the probability that the death from smoking related disease will be between 240 and 260 cases \((240 \leq X \leq 260)\)?

Answer:

\( X \), the count of deaths is \( B(1000, 0.25) \). This can be approximated by \( N(250, 13.7^2) \) since we have \( np = 250, \ nq = 750 \).

\[
P(240 < X < 260) \approx P \left( \frac{240 - 250}{13.7} < Z < \frac{260 - 250}{13.7} \right) \approx P(-0.73 < Z < 0.73)
\]

\[
= 0.5346.
\]

Example 2.

On average 2% of batteries produced by a car battery manufacturer are defective. What is the chance that in a stock of 10,000 batteries there are no more than 180 defective?
Answer:

\( X \), the count of defective batteries in a stock of 10,000 is \( B(10000,0.02) \). This can be approximated by \( N(200,196) \). Obviously \( np = 200 > 5 \), \( nq = 9800 > 5 \).

\[
P(X < 180) \approx P\left( Z < \frac{180 - 200}{14} \right) \approx P(Z < -1.43) = 0.0764.
\]

Example 3 (Small ranges or specific value of the variable).

On average one out of 6 smokers is going to die from the lung cancer. If there are 120 smokers in my neighborhood, what is the chance there will be 20 of them that will die of lung cancer?

Answer:

\( X \), the count of rounds won is \( B(120,1/6) \). This can be approximated by \( N(20, 16.67) \).

\[
P(X = 20) = P(19.5 < X < 20.5) \approx P\left( \frac{19.5 - 20}{4.08} < Z < \frac{20.5 - 20}{4.08} \right) \approx P(-0.12 < Z < 0.12)
\]

\[
\approx 0.0956
\]

One can calculate this probability directly from formula for the mass of \( B(100,1/38) \).

\[
P(X = 3) = \binom{120}{20} \left( \frac{1}{6} \right)^{20} \left( \frac{5}{6} \right)^{100} \approx 0.0973 .
\]

Obviously the error of normal approximation to binomial was less than 0.2%.

**Homework:** Check online.