Central Limit Theorem

STAT 330 - Iowa State University

Outline

In this lecture, students will be introduced to the Central Limit Theorem. They will use this theorem to calculate probabilities for sums and averages of random variables.

Central Limit Theorem (CLT)

Suppose $X_1, X_2, \ldots, X_n$ are iid random variables. For $i = 1, \ldots, n$,

$$X_i \overset{iid}{\sim} \text{distribution}$$

Any function of $\{X_i\}$ is also a random variable. Specifically,

- $S_n = \sum_{i=1}^{n} X_i$ is a R.V (with some distribution)
- $\overline{X_n} = \frac{\sum_{i=1}^{n} X_i}{n}$ is a R.V (with some distribution)

For large sample size n, the distribution of S_n and $\overline{X}$ both follow normal distributions!

Even without knowing the distribution of $\{X_i\}$, we can calculate probabilities for sums and averages using the normal distribution. (extremely useful for real life problems)!

Central Limit Theorem (CLT)

- Sums and averages of RVs from *any* distribution have approximately normal distributions for large sample sizes

Central Limit Theorem (CLT)

Suppose $X_1, X_2, \ldots, X_n$ are iid random variables with $\mathbb{E}(X_i) = \mu$ and $Var(X_i) = \sigma^2$ for $i = 1, \ldots, n$.
Define:

1. sample mean: $\overline{X_n} = \frac{\sum_{i=1}^n X_i}{n}$
2. sample sum: $S_n = \sum_{i=1}^n X_i$

Then, for *large n*,

$$\overline{X_n} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$S_n \approx N(n\mu, n\sigma^2)$$

How to Use CLT for Means

- For large n,

$$\overline{X_n} \approx N\left(\mu, \frac{\sigma^2}{n}\right)$$

- How to calculate probabilities involving $\overline{X_n}$?
- Standardize $\overline{X_n}$ to turn it into a standard normal random variable Z, and use the $z-$table!
- Standardize any normal random variable by subtracting <u>its mean</u>, and dividing by <u>its standard deviation</u>.

$$Z = \frac{\overline{X_n} - \mu}{\sigma/\sqrt{n}}$$

$$Z \sim N(0, 1)$$

How to Use CLT for Means Cont.

- Ex: $P(a < \overline{X_n} < b) = ?$
- Standardize all of the quantities involved in the above probability. Then use Z-table to obtain probabilities.

$$\begin{aligned}
P(a < \overline{X_n} < b) &= P\left(\frac{a-\mu}{\sigma/\sqrt{n}} < \frac{\overline{X_n}-\mu}{\sigma/\sqrt{n}} < \frac{b-\mu}{\sigma/\sqrt{n}}\right) \\
&= P\left(\frac{a-\mu}{\sigma/\sqrt{n}} < Z < \frac{b-\mu}{\sigma/\sqrt{n}}\right) \\
&= P\left(Z < \frac{b-\mu}{\sigma/\sqrt{n}}\right) - P\left(Z < \frac{a-\mu}{\sigma/\sqrt{n}}\right) \\
&= \Phi\left(\frac{b-\mu}{\sigma/\sqrt{n}}\right) - \Phi\left(\frac{a-\mu}{\sigma/\sqrt{n}}\right)
\end{aligned}$$

How to Use CLT for Sums

- For large n,

$$S_n \approx N(n\mu, n\sigma^2)$$

- Standardize S_n by subtracting its mean, and dividing by its standard deviation.

$$Z = \frac{S_n - n\mu}{\sqrt{n\sigma^2}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$
$$Z \sim N(0,1)$$

- Then, use the $Z-$table to obtain desired probabilities.
- Ex:

$$\begin{aligned} P(S_n < a) &= P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} < \frac{a - n\mu}{\sigma\sqrt{n}}\right) \\ &= P\left(Z < \frac{a - n\mu}{\sigma\sqrt{n}}\right) \\ &= \Phi\left(\frac{a - n\mu}{\sigma\sqrt{n}}\right) \end{aligned}$$

Examples

Examples

Example 1: The time you spend waiting for the bus each day has a uniform distribution between 2 minutes and 5 minutes. Suppose you wait for the bus every day for a month (30 days).

1. Let X_i = time spent waiting for the bus on the i^{th} day for $i = 1, \ldots, 30$.
   What is the distribution of each X_i?
   $X_i \sim \text{Unif}(2, 5)$
   What is it's expected value and variance?
   $\mathbb{E}(X_i) = \frac{5+2}{2} = 3.5$
   $Var(X_i) = \frac{(5-2)^2}{12} = .75$

Examples

2. Let $\overline{X}_n$ be the average time spent waiting for the bus per day over the month. $\overline{X}_n = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\sum_{i=1}^{30} X_i}{30}$

   What is the (approximate) probability that the average time you spent waiting for the bus per day is less than 4 min?

   We want $\mathbb{P}(\overline{X}_{30} < 4)$

   First we know, from the CLT, $\overline{X}_{30} \approx N(3.5, .025)$, so we use the normal distribution.

Examples

$$
\begin{aligned}
\mathbb{P}(\overline{X}_{30} < 4) &= \mathbb{P}\left(Z < \frac{4 - 3.5}{.158}\right) \\
&= \mathbb{P}(Z < 3.16) \\
&= \Phi(3.16) \\
&= .9992
\end{aligned}
$$

Examples

3. How much time do you expect to spend waiting for the bus in total for a month?
   Let S_{30} = total time spent waiting in the month (30 days)
   $S_{30} = \sum_{i=1}^{30} X_i$
   $\mathbb{E}(S_{30}) = 30\mu = 30 \cdot 3.5 = 105(\text{mins})$
4. What is the (approximate) probability that you spend more than 2 hours waiting for a bus in total for a month?
   Let S_{30} = total time spent waiting in the month (30 days)
   From the CLT, $S_{30} \approx N(105, 22.5)$

$$
\begin{aligned}
\mathbb{P}(S_{30} > 120) &= \mathbb{P}\left(Z > \frac{120 - 105}{\sqrt{22.5}}\right) \\
&= \mathbb{P}(Z > 3.16) \\
&= 1 - \Phi(3.16) \\
&= .0008
\end{aligned}
$$

Examples

Example 2: Suppose an image has an expected size 1 megabyte with a standard deviation of 0.5 megabytes. A disk has 330 megabytes of free space. Is this disk likely to be sufficient for 300 independent images?

Let $X_i =$ size of a random image.

We have $\mathbb{E}(X_i) = 1$ and $Var(X_i) = 0.5^2 = 0.25$

Let $S_{300} =$ total size of 300 images

$S_{300} = \sum_{i=1}^{300} X_i$, where $X_i \overset{iid}{\sim} f_X(x)$

The CLT says $S_{300} \approx N(300, 75)$

Examples

We want $\mathbb{P}(S_{300} < 330)$:

$$\begin{aligned}
\mathbb{P}(S_{300} < 330) &= \mathbb{P}\left(Z < \frac{330 - 300}{\sqrt{75}}\right) \\
&= \mathbb{P}(Z < 3.46) \\
&= \Phi(3.46) \\
&= .9997
\end{aligned}$$

Examples

Example 3: An astronomer wants to measure the distance, d, from the observatory to a star. The astronomer plans to take n measurements of the distance and use the sample mean to estimate the true distance. From past records of these measurements the astronomer knows the standard deviation of a single measurement is 2 parsecs. How many measurements should the astronomer take so that the chance that his estimate differs by d by more than 0.5 parsecs is at most 0.05?

Examples

Let X_i = the ith measurement

We assume $X_1, X_2, \ldots, X_n \overset{iid}{\sim} f_X(x)$,

Where $\mathbb{E}(X_i) = d$ and $Var(X_i) = 4$

Let $\overline{X}_n = \frac{\sum_{i=1}^n X_i}{n}$

We want to find n such that $\mathbb{P}(|\overline{X}_n - d| > .5) \leq .05$

First, from the CLT, $\overline{X}_n \approx N\left(d, \frac{4}{n}\right)$

Examples

Second, it is easier to re-write the probability as
$\mathbb{P}(|\overline{X}_n - d| < .5) \geq .95$

Thus we have:

$$\begin{aligned}
\mathbb{P}(|\overline{X}_n - d| < .5) \geq .95 &= \mathbb{P}(-.5 < \overline{X}_n - d < .5) \geq .95 \\
&= \mathbb{P}\left(\frac{-.5}{2/\sqrt{n}} < \frac{\overline{X}_n - d}{2/\sqrt{n}} < \frac{.5}{2/\sqrt{n}}\right) \geq .95 \\
&= \mathbb{P}\left(\frac{-.5}{2/\sqrt{n}} < Z < \frac{.5}{2/\sqrt{n}}\right) \geq .95
\end{aligned}$$

Examples

This implies:

$$\mathbb{P}\left(Z < \frac{.5}{2/\sqrt{n}}\right) \geq .975$$

If we do a reverse look up, we find $\Phi^{-1}(.975) = 1.96$ So:

$$\frac{.5}{2/\sqrt{n}} = 1.96 \Rightarrow n = \left(\frac{2(1.96)}{.5}\right)^2 = 61.46$$

Thus the astronomer should take $n = 62$

Recap

Students should now be familiar with the Central Limit Theorem. They should be able to use the theorem when calculating probabilities for sums or averages.