

Lecture 15

Central Limit Theorem

Manju M. Johny

STAT 330 - Iowa State University

1 / 16

Central Limit Theorem (CLT)

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

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

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)
- $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$ is a R.V (with some distribution)

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

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

Idea
of
Central
Limit
Theorem

Central Limit Theorem (CLT)

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

start w/ any
dist
ex: Exp, Unif,
normal, gamma,
etc

that has
a mean &
variance

Central Limit Theorem (CLT)

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

Define:

- sample mean: $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$
- sample sum: $S_n = \sum_{i=1}^n X_i$

Then, for large n,

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

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

get the
 μ and σ^2
from original
 x_i distribution

3 / 16

How to Use CLT for Means

- For large n,

$$\bar{X}_n \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

- How to calculate probabilities involving \bar{X}_n ?
- Standardize \bar{X}_n to turn it into a standard normal random variable Z , and use the z-table! (lecture notes 14) 15
- Standardize any normal random variable by subtracting its mean, and dividing by its standard deviation.

$$Z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

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

std-normal dist
use z-table to
obtain CDF
(probabilities)

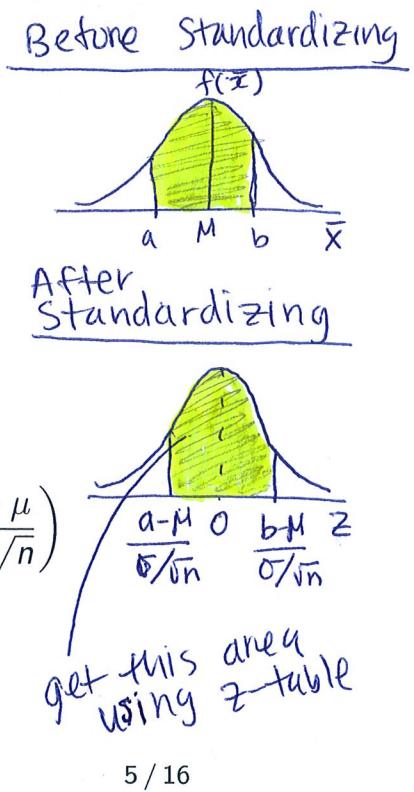
4 / 16

How to Use CLT for Means Cont.

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

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

use z-table use z-table



5 / 16

How to Use CLT for Sums

- For large n,

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

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

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

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

- Ex:

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

$\text{Ex: } \Phi(1.66)$

Look up $z = 1.66$ in margins of z-table, evaluates to just a number

and get $P(Z < 1.66)$ inside z-table

6 / 16

Examples

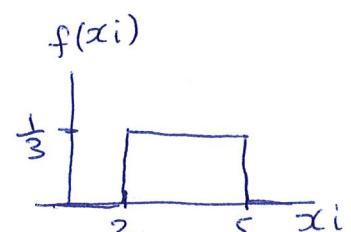
Examples

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 $i = 1, \dots, 30$.

What is the distribution of each X_i ?

For $i = 1, \dots, 30$ $\begin{cases} a=2 \\ b=5 \end{cases}$
 $X_i \sim \text{Unif}(2, 5)$



What is its expected value and variance?

For each $\begin{cases} X_i \\ (\text{each day}) \end{cases}$

$$E(X_i) = \frac{a+b}{2} = \frac{2+5}{2} = \frac{7}{2} = 3.5 = "μ"$$
$$\text{Var}(X_i) = \frac{(b-a)^2}{12} = \frac{(5-2)^2}{12} = \frac{9}{12} = 0.75 = "σ^2"$$

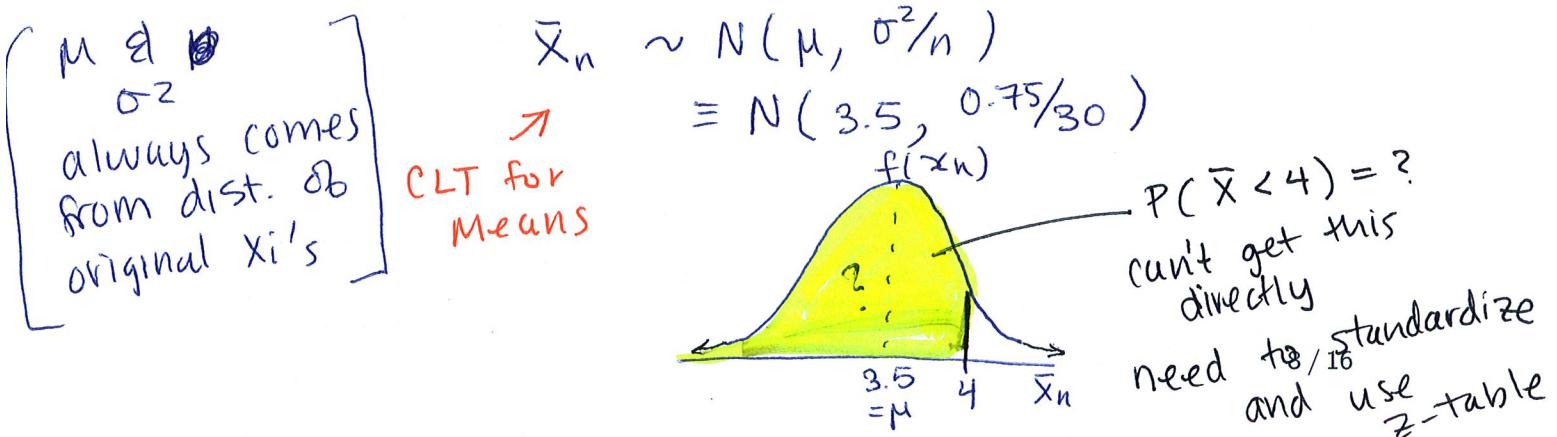
Examples

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

What is the (approximate) probability that the average time you spent waiting for the bus is less than 4 min? $P(\bar{X}_n < 4) = ?$

Now, we're interested in the R.V $\bar{X}_n = \frac{\sum_{i=1}^{30} X_i}{30}$

Since n is large,



Examples

standardize \bar{X} into R.V Z

$$Z = \frac{\bar{X}_n - M}{\sigma/\sqrt{n}} = \frac{\bar{X}_n - 3.5}{\sqrt{0.75/30}} = \frac{\bar{X}_n - 3.5}{0.1581} \sim N(0, 1)$$

$$P(\bar{X}_n < 4) = P\left(\frac{\bar{X}_n - M}{\sigma/\sqrt{n}} < \frac{4 - M}{\sigma/\sqrt{n}}\right)$$

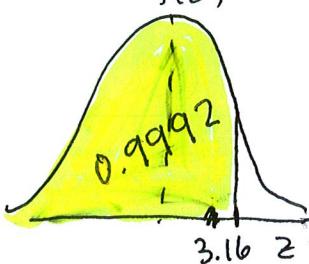
$f(z)$

$$= P\left(Z < \frac{4 - 3.5}{0.1581}\right)$$

$$= P(Z < 3.16)$$

$$= \Phi(3.16)$$

$$= 0.9992$$



use z-table

- Look up $z = 3.16$ in margins of z-table
- obtain $P(Z < 3.16)$ from inside z-table.

Examples

3. How much time do you expect to spend waiting for the bus in total for a month? Now we want $\sum_{i=1}^{30} X_i$ as our R.V
- $$E\left(\sum_{i=1}^{30} X_i\right) = 30 E(X_1) = 30 \cdot \mu = 30 \cdot 3.5 = 105$$

4. What is the (approximate) probability that you spend more than 2 hours waiting for a bus in total for a month?

120 min
our new R.V is $S_n = \sum_{i=1}^{30} X_i$

since n is large,

using CLT for sums $\rightarrow S_n \sim N(n\mu, n\sigma^2)$
 $\equiv N(30 \cdot 3.5, 30 \cdot 0.75)$
 $\equiv N(105, 22.5)$ 10 / 16

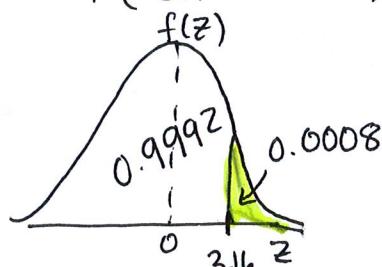
Examples

Want $P(S_n > 120) = ?$ can't get this directly
need to standardize & use z-table

standardize

$$Z = \frac{S_n - n\mu}{\sqrt{n}\sigma} = \frac{S_n - n\mu}{\sigma\sqrt{n}} = \frac{S_n - 105}{\sqrt{22.5}}$$

~~Probability~~
 $P(S_n > 120) = P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} > \frac{120 - n\mu}{\sigma\sqrt{n}}\right)$



$$= P(Z > \frac{120 - 105}{\sqrt{22.5}})$$

$$= P(Z > 3.16)$$

$$= 1 - P(Z < 3.16)$$

$$= 1 - \Phi(3.16)$$

$$= 1 - 0.9992 = 0.0008$$

Examples

single X_i

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?

$$E(X_i) = 1 = \mu \quad \text{Var}(X_i) = 0.5^2 = \sigma^2$$

We're interested in the size of the sum of 300 images

$$S_n = \sum_{i=1}^{300} X_i$$

Since $n = 300$ is large, use CLT for sums,

$$\begin{aligned} S_n &\sim N(n\mu, n\sigma^2) \\ &= N(300 \cdot 1, 300 \cdot 0.5^2) \\ &= N(300, 75) \end{aligned}$$

We want to know if 330 MB is enough space

$$\text{i.e.) } P(S_n \leq 330) = ?$$

Examples

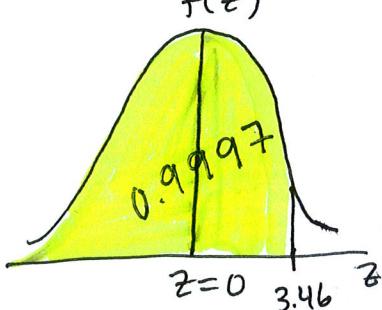
$$P(S_n \leq 330) = P\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq \frac{330 - n\mu}{\sigma\sqrt{n}}\right)$$

$$= P\left(Z \leq \frac{330 - 300}{\sqrt{75}}\right)$$

$$= P(Z \leq 3.46)$$

$$= \Phi(3.46)$$

$$= 0.9997$$



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?

For $i = 1, \dots, n$

$$X_i = \text{single measurement}$$

$$\text{Var}(X_i) = 2^2 = \sigma^2$$

$$E(X_i) = \mu = d$$

$X_i = \text{single measurement}$

$$\text{Var}(X_i) = \sigma^2 = 2^2 = 4$$

Our new R.V is $\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n}$

$$P(|\bar{X}_n - d| > 0.5) \leq 0.05$$

We want the minimum # of measurement (n) for this to be true

14 / 16

Examples

We know that $P(|\bar{X}_n - d| > 0.5) = P(\bar{X}_n - d > 0.5)$

$$+ P(\bar{X}_n - d < -0.5)$$

Using CLT for means, the distribution of \bar{X}_n is

$$\bar{X}_n \sim N(\mu, \sigma^2/n) = N(d, 4/n)$$

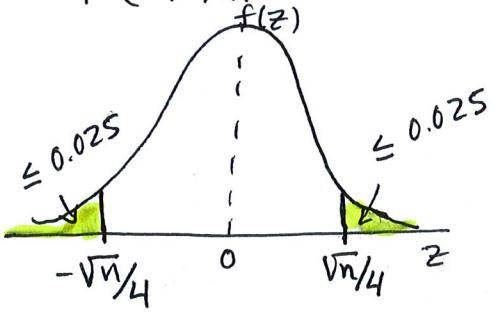
$$P(|\bar{X}_n - d| > 0.5) = P(\bar{X}_n - d > 0.5) + P(\bar{X}_n - d < -0.5)$$

$$= P\left(\frac{\bar{X}_n - d}{\sqrt{4/n}} > \frac{0.5}{\sqrt{4/n}}\right) + P\left(\frac{\bar{X}_n - d}{\sqrt{4/n}} < \frac{-0.5}{\sqrt{4/n}}\right)$$

$$= P(Z > 0.5/\sqrt{4/n}) + P(Z < -0.5/\sqrt{4/n})$$

$$= P(Z > \sqrt{n}/4) + P(Z < -\sqrt{n}/4)$$

$$= 2 \Phi(-\sqrt{n}/4)$$



We need the smallest integer n such that

$$P(|\bar{X}_n - d| > 0.5) = 2 \Phi(-\sqrt{n}/4) \leq 0.05$$

15 / 16

Examples

$$\Rightarrow 2 \Phi(-\sqrt{n}/4) \leq 0.05$$

$$\Rightarrow \Phi(-\sqrt{n}/4) \leq 0.025$$

$$\Rightarrow -\sqrt{n}/4 \leq \Phi^{-1}(0.025)$$

$$\Rightarrow -\sqrt{n}/4 \leq -1.96$$

$$\Rightarrow -\sqrt{n} \leq 4(-1.96)$$

$$\Rightarrow \sqrt{n} \geq 4(1.96)$$

$$\Rightarrow n \geq (4 \cdot 1.96)^2 = 61.47$$

- Find z in the margins s.t $P(Z \leq z) = 0.025$
- Find 0.025 inside z -table, obtain corresponding z -score from margins
 $\Rightarrow z = -1.96$

We need at least $n=62$ observations