Computer Science

CSC196: Analyzing Data

Sampling Distributions

Jason Pacheco and Cesim Erten

Review of Previous Lecture

Let $X_1, \cdots, X_n$ iid with mean μ and variance σ^2.

Probability distribution of $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ is sampling distribution of the mean.

$E(\overline{X}) =$ $\qquad$ $Var(\overline{X}) =$

Review of Previous Lecture

Let $X_1, \cdots, X_n$ iid with mean μ and variance σ^2.

Probability distribution of $\overline{X} = \dfrac{1}{n}\sum_{i=1}^{n} X_i$ is sampling distribution of the mean.

$E(\overline{X}) = \mu$ $\qquad$ $Var(\overline{X}) = \dfrac{\sigma^2}{n}$

Outline

- Sampling Distributions (continued)
- Central Limit Theorem

Examples Combining Sample Mean with Chebyshev's

Example: Want to take a random sample from a distribution with unknown μ and variance σ^2 is 8.

Size of sample such that $\overline{X}$ is within 2 units of μ with ≥ 0.99 probability?

Examples Combining Sample Mean with Chebyshev's

Example: Want to take a random sample from a distribution with unknown μ and variance σ^2 is 8.

Size of sample such that $\overline{X}$ is within 2 units of μ with ≥ 0.99 probability?

$$P(|X - \mu| \leq t) \geq 1 - \frac{\sigma^2}{t^2}$$ $\rightsquigarrow$ Statement of Chebyshev's Theorem

Examples Combining Sample Mean with Chebyshev's

Example: Want to take a random sample from a distribution with unknown μ and variance σ^2 is 8.

Size of sample such that $\overline{X}$ is within 2 units of μ with ≥ 0.99 probability?

$$P(|X-\mu| \leq t) \geq 1 - \frac{\sigma^2}{t^2}$$

$\rightsquigarrow$ statement of Chebyshev's Theorem

$$P(|\bar{X}-\mu| \leq 2) \geq 1 - \frac{1}{4} \times \frac{8}{n}$$

Examples Combining Sample Mean with Chebyshev's

Example: Want to take a random sample from a distribution with unknown μ and variance σ^2 is 8.

Size of sample such that $\overline{X}$ is within 2 units of μ with ≥ 0.99 probability?

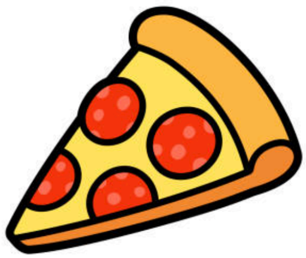
$$P(|X-\mu| \leq t) \geq 1 - \frac{\sigma^2}{t^2}$$ ⇝ Statement of Chebyshev's Theorem

$$P(|\bar{X}-\mu| \leq 2) \geq 1 - \frac{1}{4} \times \frac{8}{n}$$

$$\Rightarrow 1 - \frac{2}{n} = 0.99 \Rightarrow n = \frac{2}{0.01} = \underline{\underline{200}}$$

Central Limit Theorem

Example: Time it takes to deliver a pizza: exponential distribution



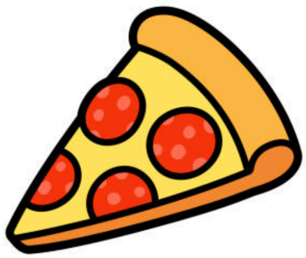
Mean delivery time: 15 minutes

Consider the average delivery times of 30 drivers over a week:

- Individual delivery times might be erratic.
- BUT average of 30 deliveries has a pattern.

Central Limit Theorem

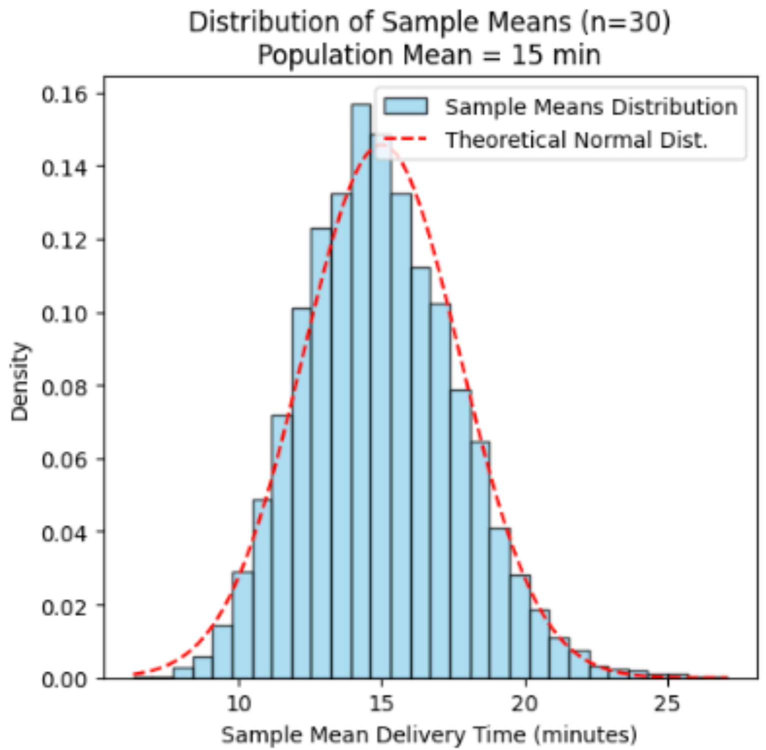
Example: Time it takes to deliver a pizza: exponential distribution



Mean delivery time: 15 minutes

Consider the average delivery times of 30 drivers over a week:

- Individual delivery times might be erratic.
- BUT average of 30 deliveries has a pattern.



Check: https://colab.research.google.com/drive/1DOHaH6GgxSQkgLQLkR1xpNBcw98lEcp4?usp=sharing

Central Limit Theorem

Informally:

For a large random sample, sample mean has an approximately normal distribution with mean μ, variance σ^2/n.

Note: Nothing is assumed regarding the distribution of X_i.

Central Limit Theorem

Informally:

For a large random sample, sample mean has an approximately normal distribution with mean μ, variance σ^2/n.

Note: Nothing is assumed regarding the distribution of X_i.

Formally: Let $X_1, \cdots, X_n$ be iid with mean μ, standard deviation σ.

$$\text{As } n \to \infty, \quad Z = \frac{\bar{X} - \mu}{\sqrt{Var(\bar{X})}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0,1)$$

Central Limit Theorem

Informally:

For a large random sample, sample mean has an approximately normal distribution with mean μ, variance σ^2/n.

Note: Nothing is assumed regarding the distribution of X_i.

Formally: Let $X_1, \cdots, X_n$ be iid with mean μ, standard deviation σ.

$$\text{As } n \to \infty, \quad Z = \frac{\bar{X} - \mu}{\sqrt{Var(\bar{X})}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0,1)$$

(Transform normal RV $x = \frac{x - \mu}{\sigma} \rightarrow z$)

Central Limit Theorem

Informally: $Var(\bar{X}) = \frac{\sigma^2}{n} \Rightarrow \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \Longrightarrow \frac{\bar{x} - \mu}{\sigma_{\bar{x}}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma}$

For a large random sample, sample mean has an approximately normal distribution with mean μ, variance σ^2/n.

Note: Nothing is assumed regarding the distribution of X_i.

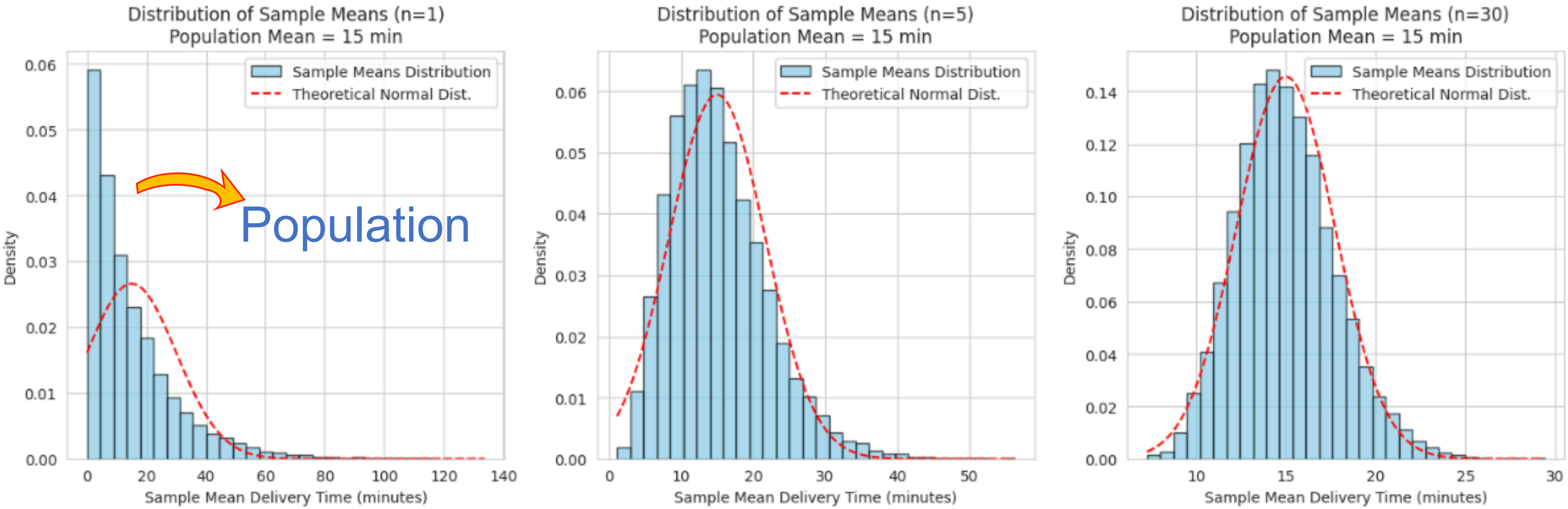
Formally: Let $X_1, \cdots, X_n$ be iid with mean μ, standard deviation σ.

$$\text{As } n \to \infty, \quad Z = \frac{\bar{X} - \mu}{\sqrt{Var(\bar{X})}} = \frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \sim N(0,1)$$

(Transform normal RV $x = \frac{x - \mu}{\sigma} \rightsquigarrow z$)

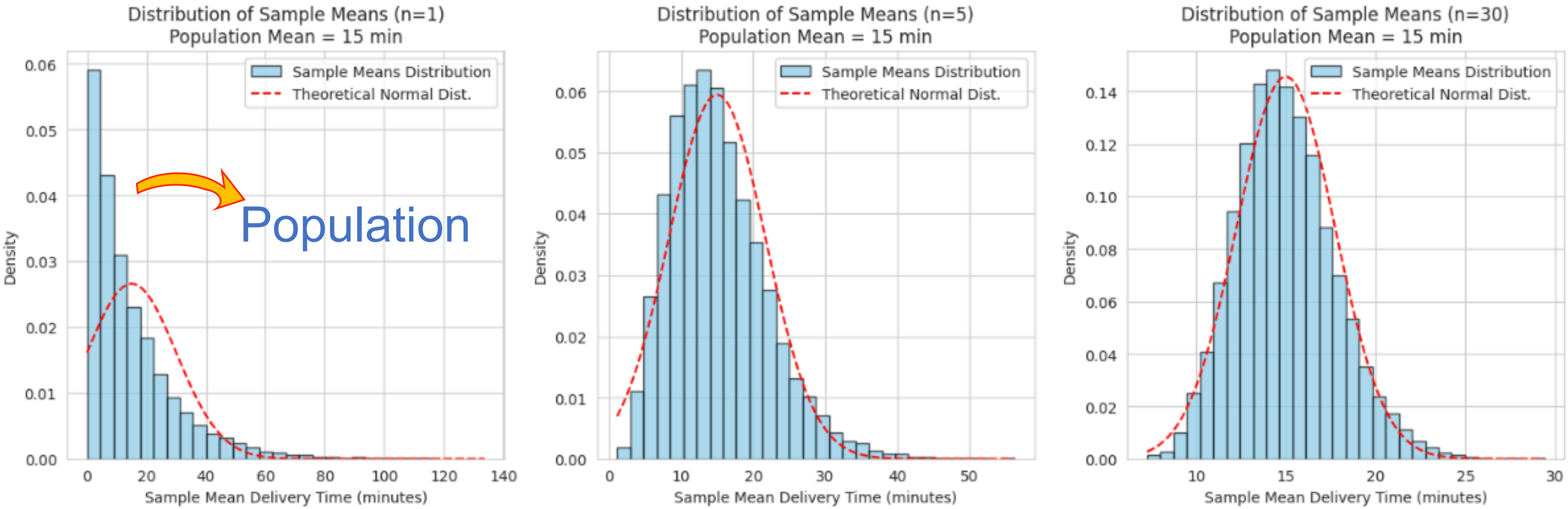
Central Limit Theorem

What sample size n? In general $n \geq 30$ should be good.



Central Limit Theorem

What sample size n? In general $n \geq 30$ should be good.



To observe the power of CLT, as an exercise: plugin a value for p in our polling example and apply the CLT.

Examples on Central Limit Theorem

Example 8.4: An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours. Use norm.cdf(-2.5)=0.0062.

Examples on Central Limit Theorem

Example 8.4: An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours. Use norm.cdf(-2.5)=0.0062.

$P(\bar{X} < 775) = ?$ [Assuming $\bar{x}$ is normal.]

Examples on Central Limit Theorem

Example 8.4: An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours. Use norm.cdf(-2.5)=0.0062.

$$P(\bar{X} < 775) = ?$$

[Assuming $\bar{x}$ is normal.]

$$P\left(\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} < \frac{775 - 800}{\frac{40}{\sqrt{16}}}\right)$$

Examples on Central Limit Theorem

Example 8.4: An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours. Use norm.cdf(-2.5)=0.0062.

$$P(\bar{X} < 775) = ?$$

[Assuming $\bar{x}$ is normal.]

$$P\left(\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} < \frac{775 - 800}{\frac{40}{\sqrt{16}}}\right) = P(Z < -2.5)$$

Z

Examples on Central Limit Theorem

Example 8.4: An electrical firm manufactures light bulbs that have a length of life that is approximately normally distributed, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours. Use norm.cdf(-2.5)=0.0062.

$P(\bar{X} < 775) = ?$ [Assuming $\bar{x}$ is normal.]

$$P\left(\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} < \frac{775 - 800}{\frac{40}{\sqrt{16}}}\right) = P(Z < -2.5)$$

$\frac{\bar{X} - \mu}{\sigma_{\bar{X}}} \to Z$

$$= 0.0062$$

Examples on Central Limit Theorem

Example 8.5: Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 4 minutes. In a given week, a bus transported passengers 36 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute. Use 1-norm.cdf(3.75)=0.000233.

Examples on Central Limit Theorem

Example 8.5: Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 4 minutes. In a given week, a bus transported passengers 36 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute. Use 1-norm.cdf(3.75)=0.000233.

$P(\bar{X} > 30.5) = ?$

Examples on Central Limit Theorem

Example 8.5: Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 4 minutes. In a given week, a bus transported passengers 36 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute. Use 1-norm.cdf(3.75)=0.000233.

$P(\bar{x} > 30.5) = ?$ $\bar{x}$ is approximately normal due to CLT.

Examples on Central Limit Theorem

Example 8.5: Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 4 minutes. In a given week, a bus transported passengers 36 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute. Use 1-norm.cdf(3.75)=0.000233.

$P(\bar{X} > 30.5) = ?$ $\bar{X}$ is approximately normal due to CLT.

$$P\left(\underbrace{\frac{\bar{X} - \mu}{\sigma_{\bar{X}}}}_{Z} > \frac{30.5 - 28}{\frac{4}{\sqrt{36}}}\right)$$

Examples on Central Limit Theorem

Example 8.5: Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 4 minutes. In a given week, a bus transported passengers 36 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute. Use 1-norm.cdf(3.75)=0.000233.

$P(\bar{X} > 30.5) = ?$ $\bar{X}$ is approximately normal due to CLT.

$$P\left(\frac{\bar{X}-\mu}{\sigma_{\bar{X}}} > \frac{30.5-28}{\frac{4}{\sqrt{36}}}\right) = P\left(Z > \frac{2.5 \times 3}{2}\right)$$

$$= P(Z > 3.75)$$

Z

Examples on Central Limit Theorem

Example 8.5: Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 4 minutes. In a given week, a bus transported passengers 36 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute. Use 1-norm.cdf(3.75)=0.000233.

$P(\bar{x} > 30.5) = ?$ $\bar{x}$ is approximately normal due to CLT.

$$P\left(\underbrace{\frac{\bar{x} - \mu}{\sigma_{\bar{x}}}}_{Z} > \frac{30.5 - 28}{\frac{4}{\sqrt{36}}}\right) = P\left(Z > \frac{2.5 \times 3}{2}\right)$$

$$= P(Z > 3.75)$$

$$= 1 - P(Z < 3.75)$$

Examples on Central Limit Theorem

Example 8.5: Traveling between two campuses of a university in a city via shuttle bus takes, on average, 28 minutes with a standard deviation of 4 minutes. In a given week, a bus transported passengers 36 times. What is the probability that the average transport time was more than 30 minutes? Assume the mean time is measured to the nearest minute. Use 1-norm.cdf(3.75)=0.000233.

$P(\bar{x} > 30.5) = ?$ $\bar{x}$ is approximately normal due to CLT.

$$P\left(\frac{\bar{x} - \mu}{\sigma_{\bar{x}}} > \frac{30.5 - 28}{\frac{4}{\sqrt{36}}}\right) = P\left(z > \frac{2.5 \times 3}{2}\right)$$

$\frac{\bar{x} - \mu}{\sigma_{\bar{x}}} \rightarrow z$

$$= P(z > 3.75)$$

$$= 1 - P(z < 3.75)$$

$$= 0.000233 \;//$$

Examples on Central Limit Theorem

Example: Number of errors per program has Poisson distribution with mean 4. Let $X_1, \cdots, X_{100}$ be the numbers of errors in 100 programs. How to approximate $P(\overline{X} < 4.2)$? Use norm.cdf(1)=0.84.

Examples on Central Limit Theorem

Example: Number of errors per program has Poisson distribution with mean 4. Let $X_1, \cdots, X_{100}$ be the numbers of errors in 100 programs. How to approximate $P(\bar{X} < 4.2)$? Use norm.cdf(1)=0.84. [Note: Variance of Poisson is the same as its mean.]

Examples on Central Limit Theorem

Example: Number of errors per program has Poisson distribution with mean 4. Let $X_1, \cdots, X_{100}$ be the numbers of errors in 100 programs. How to approximate $P(\bar{X} < 4.2)$? Use norm.cdf(1)=0.84.

[Note: Variance of Poisson is the same as its mean.]

$$P(\bar{x} < 4.2) = P\Bigg(\underbrace{\frac{\bar{x} - \mu}{\sigma_{\bar{x}}}}_{z} < \frac{4.2 - 4}{\frac{2}{\sqrt{100}}}\Bigg)$$

Examples on Central Limit Theorem

Example: Number of errors per program has Poisson distribution with mean 4. Let $X_1, \cdots, X_{100}$ be the numbers of errors in 100 programs. How to approximate $P(\bar{X} < 4.2)$? Use norm.cdf(1)=0.84.

[Note: Variance of Poisson is the same as its mean.]

$$P(\bar{X} < 4.2) = P\left(\underbrace{\frac{\bar{X} - \mu}{\sigma_{\bar{X}}}}_{Z} < \frac{4.2 - 4}{\frac{2}{\sqrt{100}}}\right)$$

$$= P\left(Z < \frac{0.2 \times 10}{2}\right) = P(Z < 1) = 0.84$$