Lesson 17: Central Limit Theorem

Meike Niederhausen and Nicky Wakim

2025-11-24

Learning Objectives

1. Calculate probability of a sample mean using a population mean and variance with unknown distribution
2. Use the Central Limit Theorem to construct the Normal approximation of the Binomial and Poisson distributions

The Central Limit Theorem

Theorem 1: Central Limit Theorem (CLT)

Let X_i be iid rv's with common mean μ and variance σ^2, for $i = 1, 2, \ldots, n$. Then

{ independent & identically distributed }

$$\sum_{i=1}^{n} X_i \to \mathrm{N}(n\mu, n\sigma^2)$$

→ converges in distribution (& probability) as $n \to \infty$

- X_i' do NOT need to be normally distributed
- don't need a known distribution of X
- in application: when n is "large", we can use Normal approx.

Extension of the CLT

Corollary 1

Let X_i be iid rv's with common mean μ and variance σ^2, for $i = 1, 2, \ldots, n$. Then

$$\overline{X} = \underbrace{\frac{\sum_{i=1}^{n} X_i}{n}} \rightarrow \underline{\mathrm{N}\left(\mu, \frac{\sigma^2}{n}\right)}$$

★ Sampling distribution of the sample means ★

BSTA 511

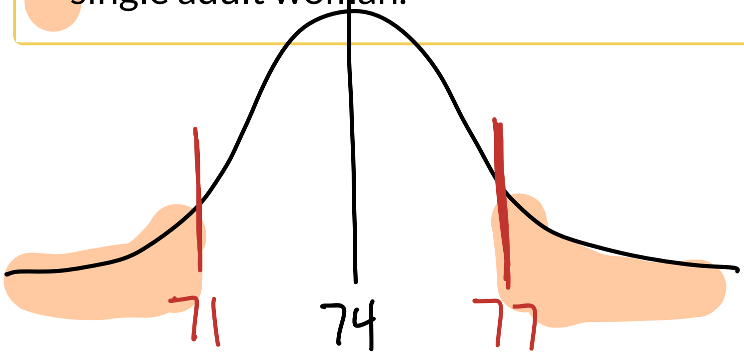
used in 512/513

Example of Corollary in use

Example 1 $\bar{X}$

According to a large US study, the mean resting heart rate of adult women is about 74 beats per minutes (bpm), with standard deviation 13 bpm (NHANES 2003-2004).

1. Find the probability that the average resting heart rate for a random sample of 36 adult women is more than 3 bpm away from the mean.
2. Repeat the previous question for a single adult woman.



$n \geq 30$ b/c 36 women is sample

$\mu = 74 \quad \sigma = 13 \quad Z^* = \mu \pm 3$
$= 71$ or 77

① $P(\bar{X} < 71$ or $\bar{X} > 77)$

$\bar{X} \sim \text{Normal}\left(\mu = 74, sd = \frac{\sigma}{\sqrt{n}} = \frac{13}{\sqrt{36}}\right)$

$P(\bar{X} < 71) + P(\bar{X} > 77)$

$= 2 \cdot P(\bar{X} < 71)$

$= 2 \cdot$ pnorm(x=71, mean=74, sd= 13/sqrt(36))

$= 0.164$

② $n = 1 < 30$, no normal approx

Rule of thumb: $n \geq 30$ to use CLT

Example of CLT for exponential distribution

Example 2

Let $X_i \sim Exp(\lambda)$ be iid RVs for $i = 1, 2, \ldots, n$. Then

$$\sum_{i=1}^{n} X_i \to$$

$$\mu = \frac{1}{\lambda}$$

$$\sigma^2 = 1/\lambda^2$$

$$\sum_{i=1}^{n} X_i \longrightarrow \text{Normal}(n\mu, n\sigma^2)$$

converges in distribution as $n \to \infty$

$$n\mu = \frac{n}{\lambda} \qquad n\sigma^2 = \frac{n}{\lambda^2}$$

$$\sum_{i=1}^{n} X_i \longrightarrow N\left(\frac{n}{\lambda}, \frac{n}{\lambda^2}\right)$$

CLT for Discrete RVs

1. Binomial rv's: Let $X \sim Bin(n, p)$

- $X = \sum_{i=1}^{n} X_i$, where X_i are iid $\text{Bernoulli}(p)$
- Rule of thumb: $np \geq 10$ and $n(1-p) \geq 10$ to use Normal approximation

$$\underbrace{\sum_{i=1}^{n} X_i}_{\text{sum of Bern is a Binomial}} \longrightarrow N(np, np(1-p))$$

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n} \longrightarrow N\left(p, \frac{p(1-p)}{n}\right)$$

2. Poisson rv's: Let $X \sim Poisson(\lambda)$

- $X = \sum_{i=1}^{n} X_i$, where X_i are iid $\text{Poiss}(1)$
- Recall from ~~Chapter 18~~ Lesson 15 that if $X_i \sim Poiss(\lambda_i)$ and X_i independent, then $\underline{\sum_{i=1}^{n} X_i} \sim Poiss(\sum_{i=1}^{n} \lambda_i)$
- Rule of thumb: $\lambda \geq 10$ to use Normal approximation

$$\sum_{i=1}^{n} X_i \longrightarrow N(\underbrace{\lambda}_{\text{mean}}, \underbrace{\lambda^2}_{\text{var}})$$

Larger example (1/7)

Example 3

Suppose that the probability of developing a specific type of breast cancer in women aged 40-49 is 0.001. Assume the occurrences of cancer are independent. Suppose you have data from a random sample of 20,000 women aged 40-49.

1. How many of the 20,000 women would you expect to develop this type of breast cancer, and what is the standard deviation?
2. Find the **exact** probability that more than 15 of the 20,000 women will develop this type of breast cancer.
3. Use the CLT to find the **approximate** probability that more than 15 of the 20,000 women will develop this type of breast cancer.
4. Use the CLT to approximate the following probabilities, where X is the number of women that will develop this type of breast cancer.
   a. $\mathbb{P}(15 \leq X \leq 22)$
   b. $\mathbb{P}(X > 20)$
   c. $\mathbb{P}(X < 20)$
5. Find the **approximate** probability that more than 15 of the 20,000 women will develop this type of breast cancer - not using the CLT!
6. Use the CLT to approximate the approximate probability in the previous question!

Larger example (2/7)

Example 3

Suppose that the probability of developing a specific type of breast cancer in women aged 40-49 is 0.001. Assume the occurrences of cancer are independent. Suppose you have data from a random sample of 20,000 women aged 40-49.

1. How many of the 20,000 women would you expect to develop this type of breast cancer, and what is the standard deviation?

let X = # women that develop type of breast cancer

$$X \sim \text{Bin}(n = 20{,}000,\ p = 0.001)$$

$$\mu = np = 20000 \cdot 0.001 = 20 \quad \text{exp val}$$

$$\sigma = \sqrt{np(1-p)} = \sqrt{20000 \cdot 0.001\,(1-0.001)}$$

$$= 4.4699 \quad \text{std dev}$$

Larger example (3/7)

Example 3

Suppose that the probability of developing a specific type of breast cancer in women aged 40-49 is 0.001. Assume the occurrences of cancer are independent. Suppose you have data from a random sample of 20,000 women aged 40-49.

2. Find the **exact** probability that more than 15 of the 20,000 women will develop this type of breast cancer.

$$P(X > 15) = \sum_{x=16}^{20000} \binom{20000}{x} 0.001^{x} \, 0.999^{20000-x}$$

or ≥ 16

$= \text{pbinom}(q = 15,\ n = 20000,\ p = 0.001,\ \underbrace{\text{lower.tail} = F}_{P(X > 15)})$

true: $P(X \leq 15)$

$= 0.843616$

Larger example (4/7)

CONT CORR

$P(\underline{X} \leq k) = P(X \leq k + 0.5)$

$X \rightarrow \underline{N}$

$P(X \geq k) = P(X \geq k - 0.5)$

Example 3

Suppose that the probability of developing a specific type of breast cancer in women aged 40-49 is 0.001. Assume the occurrences of cancer are independent. Suppose you have data from a random sample of 20,000 women aged 40-49.

3. Use the CLT to find the **approximate** probability that more than 15 of the 20,000 women will develop this type of breast cancer.

$X \longrightarrow N(\mu = np, \sigma^2 = np(1-p))$

$N(20, 19.98)$

☆ check

$np \geq 10$ $\quad$ $np(1-p) \geq 10$

$20 \geq 10$ ✓ $\quad$ $19.98 \geq 10$ ✓

so we can use CLT

$P(X > 15) = P(X \geq 15.5)$

or $P(X \geq 16)$

= pnorm(q = 15.5, mean = 20, sd = 4.4699, lower.tail = F)

= 0.8438

Normal approx not so useful in regression

Larger example (5/7)

Example 3

Suppose that the probability of developing a specific type of breast cancer in women aged 40-49 is 0.001. Assume the occurrences of cancer are independent. Suppose you have data from a random sample of 20,000 women aged 40-49.

4. Use the CLT to approximate the following probabilities, where X is the number of women that will develop this type of breast cancer.

   a. $\mathbb{P}(15 \leq X \leq 22)$
   b. $\mathbb{P}(X > 20)$
   c. $\mathbb{P}(X < 20)$

Larger example (6/7)

Example 3

Suppose that the probability of developing a specific type of breast cancer in women aged 40-49 is 0.001. Assume the occurrences of cancer are independent. Suppose you have data from a random sample of 20,000 women aged 40-49.

5. Find the **approximate** probability that more than 15 of the 20,000 women will develop this type of breast cancer - not using the CLT!

Poisson approx of Binomial

$\frac{1}{10} \leq np(1-p) \cancel{\leq} 10$

19.98

does NOT meet requirement for Poisson approx

if it met assumptions:

$Y \sim Poiss(\lambda = np = 20)$

$$P(Y > 15) = \sum_{y=16}^{\infty} \frac{e^{-20} 20^y}{y!}$$

$Y \geq 16$

$$P(Y=y) = \frac{\lambda^y e^{-\lambda}}{y!}$$

$= 0.8434869$

Approximation of Binomial → $\mu = np$

$\sigma = \sqrt{np(1-p)}$

$\mu \not\perp \sigma$

<u>Normal</u> approx

$\mu \perp \sigma$

Linear models
(assumes Normal
dist)

assume $\sigma \perp \mu$

vs.

Poisson approx

$\mu \not\perp \sigma$

Logistic regression
or Poisson regress

Larger example (7/7)

Example 3

Suppose that the probability of developing a specific type of breast cancer in women aged 40-49 is 0.001. Assume the occurrences of cancer are independent. Suppose you have data from a random sample of 20,000 women aged 40-49.

6. Use the CLT to approximate the approximate probability in the previous question!

$Y \sim Poiss(\lambda = 20) \qquad Y \xrightarrow{p} N(\mu = \lambda, \sigma^2 = \lambda)$