4.3 CENTRAL LIMIT THEORY

WK 7 - 2

MOST IMPORTANT THEORY IN STATS! IT SAYS! ALMOST EVERYTHING EVENTUALLY BECOMES NORMAL!

STANDARD DEFINITION:

1 SUPPOSE THAT THE RANDOM VARIABLES $X_1, X_2, X_3 \dots X_n$ ARE DRAWN FROM A LARGER POPULATION INDEPENDENTLY (therefore $X_1, X_2, X_3 \dots X_n$ are identically distributed) WITH a common mean μ and the standard deviation σ. LET

WE GET 3 PIECES OF INFO →

SAMPLE TOTAL ① $T_0 = X_1 + X_2 + \dots + X_n$ ↗ have identical distributions

REQUIREMENT: LARGE SAMPLE SIZE!

THEN ② EXPECTED VALUE $E(T_0) = n\mu$ and ③ SD $\sigma_{T_0} = \sigma\sqrt{n}$ (standard error)

↑ if you make a random variable from them, it would be a normal distribution

furthermore, if n is large (≥ 30 or ≥ 40), then we may closely approximate the distribution of T_0 by a normal distribution.

→ WE DON'T SELECT SAMPLE W/ REPLACEMENT, BUT FOR THE THEORY, THEY NEED TO BE INDEPENDENT. SO! WE NEED LARGE SAMPLE SIZE THAT THE PROBABILITY OF REPEATING IS LOW SO WE CAN SELECT SAMPLE W/ REPLACEMENT.

- ↳ MAGICALLY, THE THEOREM PROVES $X_1 + X_2 + \dots X_n$ WILL BE NORMAL
- ◦ but application really is case by case. If data is skewed, you would need even larger sample sizes
- ◦ IRL IF YOU DONT KNOW COMMON μ + σ YOU CAN EVEN USE THE AVERAGE μ + σ

WAIT.. WE HAVE 4 CLT'S ∽ THERE ARE 4 CASES

2 SUPPOSE THAT THE RANDOM VARIABLES $X_1, X_2, \dots X_n$ are drawn with replacement from a large population independent + identically distributed with COMMON MEAN μ + STANDARD DEVIATION σ. Let:

PAY ATTENTION TO TOTAL / AVERAGE IN HW

basically derived from SAMPLE TOTAL

SAMPLE AVERAGE (MEAN) $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n}$ THEN → $E(\bar{X}) = \mu$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$ (standard error)

furthermore, if n is large (≥ 30 or ≥ 40) then we may closely approximate the distribution of $\bar{X}$ by a normal distribution

3 A SPECIAL CASE OF 1

SUPPOSE THAT THE RANDOM VARIABLES $X_1, X_2, \dots X_n$ are drawn with replacement (↗ identical). SUPPOSE ALSO THAT $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$

ADDING BERNOULLIES! → BINOMIAL DIST.!

1 = SUCCESS 0 = FAILURE

SAMPLE TOTAL COUNT TOTAL SUCCESS

THEN X HAS A BINOMIAL DISTRIBUTION WITH $E(X) = np$ and $\sigma_X = \sqrt{np(1-p)}$ (↗ also called standard error)

FURTHERMORE, IF n IS LARGE ($np \geq 10$ and $n(1-p) \geq 10$), then we may approximate closely the distribution of X by a normal distribution

THIS APPROXIMATES A BINOMIAL DISTRIBUTION INTO A NORMAL ONE (BUT IT'S NOT EXACT ANSWER)

BUT IN HW DON'T USE NORMAL TO APPROXIMATE BINOMIAL

4 WE HAVE 3 TO GET THE PERCENTAGE!

SUPPOSE THAT THE RANDOM VARIABLES $X_1, X_2, \dots X_n$ are drawn with replacement. SUPPOSE ALSO THAT $P(X_i = 1) = p$ and $P(X_i = 0) = 1 - p$

1 = SUCCESS 0 = FAILURE

SAMPLE PERCENTAGE (PROPORTION) $\hat{P} = \frac{x}{n} = \frac{X_1 + X_2 + X_3 + \dots + X_n}{n}$ (↗ this is not binomial, you have to use CLT to get percentage)

THEN → $E(\hat{P}) = p$ + $\sigma_{\hat{P}} = \sqrt{\frac{p(1-p)}{n}}$ ← STANDARD ERROR

FURTHERMORE, IF n IS LARGE ($np \geq 10$ and $n(1-p) \geq 10$), then we may approximate closely the distribution of $\hat{P}$ by a normal distribution

ALL THE CLT'S

SAMPLE TOTAL

$E(T_0) = n\mu$

$\sigma_{T_0} = \sigma\sqrt{n}$

SAMPLE AVERAGE

$E(\bar{X}) = \mu$

$\sigma_{\bar{X}} = \sigma/\sqrt{n}$

FOLLOWS NORMAL DISTRIBUTION IF N IS LARGE

TOTAL SUCCESS

$E(X) = np$

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

↓

ACTUALLY BINOMIAL

(DON'T USE NORMAL APPROX. WHEN COMPUTING PROBABILITY)

SAMPLE PERCENTAGE

$E(\hat{p}) = p$

$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$

EXAMPLE 1: FINDING TOTAL APPROXIMATION

THIRTY-SIX JETS WAIT TO TAKE OFF FROM AN AIRPORT. THE AVERAGE TAXI + TAKE-OFF TIME FOR EACH JET IS 8.5 MINUTES, WITH AN SD OF 2.5 MINUTES. WHAT IS THE PROBABILITY THAT THE TOTAL TAXI + TAKE OFF TIME FOR THE 36 JETS IS LESS THAN 320 MIN?

WHAT IS THE VARIABLE? TIME

WHAT DISTRIBUTION DOES IT FOLLOW? WE DON'T KNOW, BUT WE CAN FIND TOTAL

IT IS CONTINUOUS

↓

TOTAL = SAMPLE SUM

$T_0 = X_1 + X_2 + \dots X_{36}$ (time per jet)

V_{ar} = time (taxi + take off)

$\mu = 8.5 \quad \sigma = 2.5 \quad n = 36$

$E(T_0) = \mu_{T_0} = 8.5 * 36$

$= 306$

$SD(T_0) = \sigma_{T_0} = \sqrt{n}\,\sigma = \sqrt{36} \cdot 2.5$

$= 15$

WE KNOW TOTAL HAS NORMAL DISTRIBUTION

$P(T_0 < 320) = \text{normalcdf}(-1000, 320, 306, 15) \approx 0.8247$

EXAMPLE 2:

THE COOKIE MACHINE AT CHIPS AHOY ADDS A RANDOM # OF CHIPS TO EACH COOKIE. THE NUMBER OF CHIPS IS A RANDOM # WITH AVERAGE 28.5 AND SD 5.3. FIND THE PROBABILITY THAT, IN A BAG OF 50 COOKIES, THE AVERAGE # OF CHIPS PER COOKIE IS AT LEAST 30.

WHAT IS THE VARIABLE? # OF CHIPS

WHAT DISTRIBUTION DOES IT FOLLOW? WE DON'T KNOW

IT IS DISCRETE THO.

$\mu = 28.5$

$\sigma = 5.3$

$n = 50$ (each cookie is a sample)

$E(\bar{x}) = \mu = 28.5$

$SD = \sigma/\sqrt{n} = 5.3/\sqrt{50}$

$\bar{x} \sim \text{NORMAL}(28.5, 5.3/\sqrt{50})$

$P(\bar{x} \geq 30) = \text{NORMALCDF}(30, 1000, 28.5, 5.3/\sqrt{50})$

(PUT IN EXACT, NO ROUNDING ERROR)

$= 0.0227$

EXAMPLE 3: BINOMIAL SUCCESS

IT IS KNOWN THAT 40% OF PEOPLE IN THE CITY ARE INFECTED BY COVID-19 (SOME NO, SOME MILD, SOME SEVERE SYMPTOMS).

A RANDOM SAMPLE OF 50 PEOPLE ARE SELECTED.

a) WHAT IS THE PROBABILITY THAT BETWEEN 200 + 210 ARE INFECTED? (USE BINOMIAL THEN NORMAL APPROX. TO SEE HOW ACCURATE)

b) WHAT IS THE PROBABILITY THAT THE SAMPLE PORTION IS BETWEEN 39% AND 43%?