01/17/2021

HEY THERE! SPRING 2021 LECTURE NOTES FOR EC362 WILL BE ON THIS DOCUMENT, PLEASE GET USED TO MY TERRIBLE HAND WRITING

01/21/2021

→ EXAMS AFTER THE SURVEY

LESSON PLAN:

- SELF-INTRODUCTION
- ECONOMETRICS: . ECON → ECONOMIC THEORY
  . METRICS → TEST IT WITH DATA.
- COURSE LOGISTICS +
- EXPECTATIONS +
- R + R STUDIO + GIT HUB → R SHOW CASE HOMEWORK ASSIGNMENT

→ RAP TALK

WHY DO WE NEED THESE?

BUSINESSES ARE CHANGING

WHAT HAPPENED TO EXCEL? → NOT TABLES.

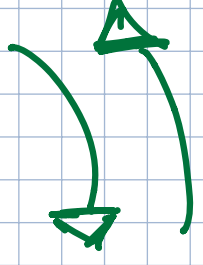
→ MARKET MORE DATA DRIVEN

→ DATA COLLECTION (MOBILE DEVICES)

STORING (CLOUD)

MANAGE (SQL, ETC)

→ MACHINE ALGORITHMS BECAME SMARTER



EX. NATURAL LANGUAGE PROCESSING.

- PREDICTION → WHAT PREDICTS SALES?
- CAUSAL CHAIN → DOES SOCIAL MEDIA PRESENCE INCREASE SALES?

WHAT SHOULD A BUSINESS STUDENT LEARN?

↳ BOTH AT AN OK LEVEL.
(UNDERSTAND + BASELINE KNOWLEDGE) → MECHANICS OF REGRESSION ANALYSIS & DATA

↳ BEST WAYS TO INFER CAUSALITY (RANDOMIZED EXPERIMENTS)

01/25/2020 LECTURE 1

LESSON PLAN:

- REVIEW OF PROBABILITY – CH 2
  - RANDOM VARIABLES
  - PROBABILITY DISTRIBUTIONS
  - E[X], VAR[X] → σ_X^2, σ_X
    - σ_X^2 ↓ VAR(X)
    - σ_X ↓ STD. DEV (X)
  - STANDARDIZED RANDOM VAR.
- APPLICATION IN R

RANDOM VARIABLE: NUMERICAL SUMMARY OF AN EVENT WITH A RANDOM COMPONENT

→ FLIPPING A COIN

→ YOUR WAGE AT AGE 40

- DISCRETE {0, 1} → SUM OF PROBABILITIES FOR ALL THE POSSIBLE OUTCOMES = 1
- CONTINUOUS {TIME} ↓ INTERVALS

EXPECTED VALUE → LONG RUN AVERAGE

RANDOM VARIABLE Y → μ_Y } EXPECTED VALUE
$E[Y]$ }

$E[Y] = Y_1 \cdot P_1 + Y_2 \cdot P_2 + \ldots + Y_k P_k$

k OUTCOMES ↓ DISCRETE PROBABILITIES

IF THE OUTCOME IS BINARY

→ DUMMY - BERNOULLI RANDOM VAR

IN GENERAL

$Y = \{0, 1\}$

$P(Y=0) = (1-P)$

$P(Y=1) = P$

$E[Y] = 0 \times (1-P) + 1 \times P$

$= P$

SURVIVE ↔ 1

COVID: MORTALITY → Y → $\{0, 1\}$

$PR(Y=1) = 0.005$ ↓ DIE

$$\mu_Y = 0 \times 0.995 + 1 \times 0.005$$
$$= 0.005$$

VARIANCE: $VAR(Y) = \sigma_Y^2$ — MEASURES THE SPREAD OF POSSIBLE OUTCOMES AROUND THE EXPECTED VALUE

$$\sigma_Y^2 = (Y_1 - \mu_Y)^2 P_1 + (Y_2 - \mu_Y) P_2 + \dots (Y_k - \mu_Y)^2 P_k$$

$$\sigma_Y^2 = \sum_i (Y_i - \mu_Y)^2 P_i$$

$\sigma_Y^2 \geq 0$ $\quad \sigma_Y = \sqrt{\sigma_Y^2}$

IF Y IS BINARY

$PR(Y=1) = P$
$PR(Y=0) = (1-P)$ $\quad \mu_Y = P$

$$VAR(Y) = (0-P)^2 \times (1-P) + (1-P)^2 \cdot P$$
$$= P(1-P)\left[P - (1-P)\right]$$
$$= P(1-P)$$

COVID MORTALITY: $VAR(Y) = 0.005 \times 0.995$

$\sigma_Y^2 = 0.004975$

$\sigma_Y \cong 0.0705$

p → CASE FATALITY RATIO

TWO RANDOM VARIABLES

$PR(Y=y)$ → PROBABILITY THAT A RANDOM VARIABLE Y (↓ CAPITAL) IS EQUAL TO A SPECIFIC VALUE y (↳ LOWER CASE)

$PR(Y=1)$ → MORTALITY RISK (IN THE PREVIOUS EXAMPLE)

JOINT PROBABILITY DISTRIBUTION OF X & Y

$PR(X=x, Y=y)$

↳ PROB. THAT $X=x$ AND $Y=y$

$X = \{0, 1\}$ → SENIOR VS. NON SENIOR

$Y = \{0, 1\}$ → MORTALITY

$PR(X=0, Y=1)$

$PR(X=1, Y=1)$

$$PR(Y=y) = \sum_{i=1}^{l} Pr(X=x_i, Y=y)$$

↳ MARGINAL PR.

↳ JOINT PROBABILITY

Ex:

TABLE 2.2 Joint Distribution of Weather Conditions and Commuting Times

	Rain ($X = 0$)	No Rain ($X = 1$)	Total
Long commute ($Y = 0$)	0.15	0.07	0.22
Short commute ($Y = 1$)	0.15	0.63	0.78
Total	0.30	0.70	1.00

$PR(X=0, Y=0) = 0.15$

$PR(X=0, Y=1) = 0.15$

$PR(X=1, Y=0) = 0.07$

$PR(X=1, Y=1) = 0.63$

} MUTUALLY EXCLUSIVE

$$PR(Y=1) = \sum_{i=1}^{2} PR(X=x_i, Y=1)$$

$$= PR(X=0, Y=1) + PR(X=1, Y=1)$$

$= 0.78$ } 1

$PR(Y=0) = 0.22$

$PR(X=0) = 0.30$ } 1

$PR(X=1) = 0.70$

BAYES THEOREM

$$PR(Y=y \mid X=x) = \frac{PR(X=x, Y=y)}{PR(X=x)}$$

$$PR(\text{SHORT COMMUTE} \mid \text{RAIN})$$

$$PR(Y=1 \mid X=0) = \frac{PR(Y=1, X=0)}{PR(X=0)}$$

$$\text{LO} \quad \frac{0.15}{0.30} = 50\%$$

01/28/2021

LESSON PLAN:

- REVIEW JOINT PROB.
- CONDITIONAL EXPECTED VALUE
- LAW OF ITERATED EXPECTATIONS
- INDEPENDENCE, COVARIANCE, CORRELATION
- SOLVE THE HANDOUT QUESTIONS

CHAPTER 2

MAY BE: INTRODUCE RANDOM SAMPLING

TWO RANDOM VARIABLES → JOINT PROBABILITY DISTRIBUTION

$PR(X=x, Y=y)$ → NOTATION for JOINT PROBABILITY

$$PR(Y=y) = \sum_{i=1}^{l} PR(X=x_i, Y=y)$$

BAYES THEOREM:

$$PR(Y=y \mid X=x) = \frac{PR(X=x, Y=y)}{PR(X=x)}$$

$$E[Y] = \sum_i y_i \, PR(Y=y_i)$$

$$E[Y \mid X=x] = \sum_i y_i \, PR(Y=y_i \mid X=x)$$

ROLL A DIE → OUTCOME Y

→ IF EVEN 0
IF ODD 1
} X

$$E[Y] = \sum_i y_i \, PR(Y=y_i)$$

$$= 1 \times 1/6 + 2 \times 1/6 + \ldots + 6 \times 1/6$$

$$= 3.5$$

$$E[Y|X=1] = \sum_i y_i \, PR(Y_i = y_i \mid X=1)$$

$$= 1 \times 1/3 + 2 \times 0 + 3 \times 1/3 + 0 + 5 \times 1/3 + 0$$

$$E[Y|X=1] = 3$$

LAW OF ITERATED EXPECTATIONS

$$E[Y] = \sum_{i=1}^{l} E[Y|X=x_i] \, PR(X=x_i)$$

$$E[Y] = E[Y|X=0] \, PR(X=0) + E[Y|X=1] \, PR(X=1)$$

$$E[Y|X=1] = 1 \times 0 + 2 \, 1/3 + 3 \times 0 + 4 \, 1/3 + 0 + 6 \, 1/3$$

$$= 4$$

$$E[Y] = 4 \times 1/2 + 3 \times 1/2$$

$$= 3.5$$

$E[Y] = E[E[Y|X]]$ → SHORT NOTATION FOR LAW OF (ITERATED) EXPECTATIONS

INDEPENDENCE: $Pr(Y=y|X=x) = Pr(Y=y)$

↳ KNOWING X PROVIDES NO INFORMATION ON THE PROBABLE OUTCOME OF Y

$$Pr(Y=Y|X=x) = \frac{Pr(X=x, Y=y)}{Pr(X=x)}$$

$$Pr(Y=y)\,Pr(X=x) = Pr(X=x, Y=y)$$

ONLY HOLDS IF X & Y ARE INDEPENDENT

$$cov(x,y) = \sigma_{xy} = E[(x-\mu_x)(Y-\mu_Y)]$$

$$= \Sigma_i \Sigma_j (x_j-\mu_x)(Y_i-\mu_i)\,Pr(X=x_i, Y=y_i)$$

$$cor(X,Y) = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

$<0, =0, >0$

ONLY INDICATES DIRECTION

$$-1 \leq cor(X,Y) \leq 1$$

↓ PERFECT NEGATIVE COR.

0 ↓ NO CORRELATION UNCORRELATED

↳ PERFECT POSITIVE COR.

QUESTION 3.

$PR(Y=y) = \sum_i PR(X=x_i, Y=y)$

$PR(A=0) = PR(A=0, M=0) + PR(A=0, M=1) + \dots$

$= 70\%$

$PR(A=1) = 50\%$

$$PR(M=0 \mid A=0) = \frac{PR(M=0, A=0)}{PR(A=0)}$$

$$= 0.35 / 0.5 = 70\%$$

$E[M \mid A=0] = \sum_i m_i \, PR(M=m \mid A=0)$

$0 \times 0.70 + 1 \times 0.13 + \dots = 0.56$

$0 \times 0.90 + 1 \times 0.07 + \dots = 0.14$

$E[M] = E[E[M \mid A]]$

$= E[M \mid A=0] \times PR(A=0) +$

$E[M \mid A=1] \times PR(A=1)$

$= 0.56 \times 0.5 + 0.14 \times 0.5 = 0.35 = E[M]$

2/1/2021 - LESSON PLAN - CH8 REVIEW OF STATS

- RANDOM SAMPLING AND DISTRIBUTION OF THE SAMPLE MEAN
- NORMAL DISTRIBUTION

POPULATION DATA → DOES NOT EXIST (IN MOST CASES)

(a) μ_x, $VAR(x)$ → ALL UNKNOWN

$E[x]$ σ_x^2 σ_x STD.DEV

COV. COR → NOT OBSERVED

ρ_{xy} σ_{xy}

RANDOM SAMPLING → n OBJECTS RANDOMLY DRAWN FROM THE POPULATION (INDIVIDUALS)

Y → IQ → EXAMPLE

↓ RANDOM SAMPLE OF n

$\{Y_1, Y_2, Y_3, \ldots, Y_n\}$ →

- ALL Y ARE DRAWN FROM THE SAME POP.
- IDENTICALLY & INDEPENDENTLY DISTRIBUTED
- i.i.d → KNOWING Y_1 PROVIDES NO INFO ON Y_2.

$P(Y_1 \le y) = P(Y_2 \le y)$

Y_1 & Y_2 ARE IDENTICALLY DISTRIBUTED.

KNOWING FIRST COIN PROVIDES NO INFORMATION ON THE SECOND

Y_1 → HEADS

Y_2 → ?

$Y_1, Y_2, \ldots, Y_N$ IID → THROUGH RANDOM SAMPLING

↳ HAVE TO BE DONE THROUGH MACHINE ALGORITH

$\bar{Y} = \sum_i Y_i / n$

$= (Y_1 + Y_2 + \ldots + Y_N) / n$

↳ USUALLY PROVIDES THE BEST INFO INSTEAD OF MEDIAN, Y_1, Y_{10}

$n = 10{,}000$

↳ ASSUME

$\bar{Y}$ → RANDOM VARIABLE

WHY?

↳ EACH 10,000 RANDOM SAMPLE WILL GIVE YOU A DIFFERENT $\bar{Y}$

$\bar{Y}$ → HAS A RANDOM COMPONENT WITH AN EXPECTED VALUE AND PROBABILITY DISTRIBUTION.

$Y \rightarrow E[Y] = \mu_Y$ $\quad VAR(Y) = \sigma_Y^2$

$\bar{Y} \rightarrow E[\bar{Y}] = \mu_Y$ $\quad VAR(\bar{Y}) = \sigma_Y^2 / n$

SEE THE TEXTBOOK FOR PROOF (APPENDIX) ↔ $\sigma_{\bar{Y}}^2 = \sigma_Y^2 / n$

$Y \to E[Y] = \mu_Y$

$\bar{Y} \to E[\bar{Y}] = \mu_Y$

$Y \to VAR(Y) \to \sigma_Y^2$

$\bar{Y} \to VAR(\bar{Y}) \to \sigma_{\bar{Y}}^2 \to \sigma_Y^2/n$

$Y \to STD.DEV(Y) \to \sigma_Y$

$\bar{Y} \to STD.DEV(\bar{Y}) \to \sigma_{\bar{Y}} \to \sigma_Y/\sqrt{n}$

PROBLEM IS STILL THERE → WE DON'T KNOW μ_Y, σ_Y^2

LARGE SAMPLE APPROXIMATIONS

UNDER GENERAL CONDITIONS $\bar{Y}$ IS AS GOOD AS μ_Y

$\bar{Y} \xrightarrow{P} \mu_Y$ → LAW OF LARGE NUMBERS

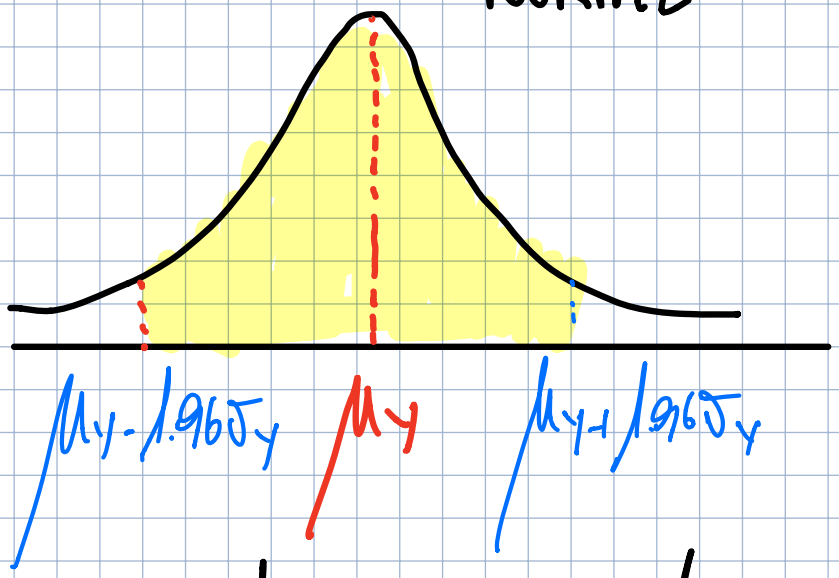
1. $Y_1, Y_2, \ldots, Y_n$ ARE iid
   WITH $E[Y_i] = \mu_Y$ → ALL DRAWS FROM SAME DIST. → NO FREQUENT BIG OUTLIERS
2. $VAR(Y_i) = \sigma_Y^2 < \infty$ → FINITE VARIANCE

NORMAL DISTRIBUTION: $Y \sim N(\mu_Y, \sigma_Y^2)$

$Y \sim N(\mu_Y, \sigma_Y^2)$ → NORMAL

$Z \sim N(0,1)$ → STANDARD NORMAL

USUALLY BOTH ARE UNKNOWN (WE WILL PROXY THESE WITH SAMPLE METRICS)



→ SYMETRIC

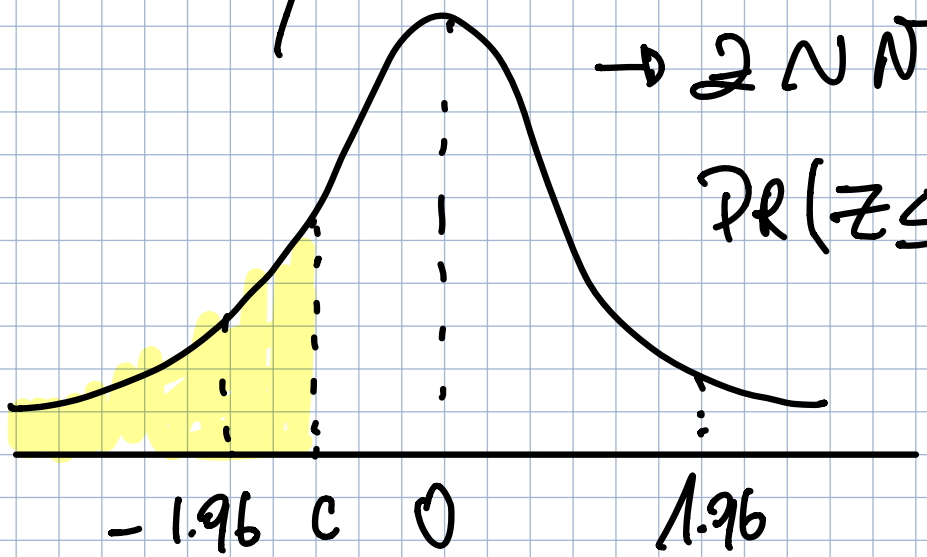
→ TAILS ASYMTOTIC

→ $(-\infty, \infty)$

$$PR(\mu_Y - 1.96\sigma_Y < Y < \mu_Y + 1.96\sigma_Y) = 95\%$$

$$PR(Y < \mu_Y - 1.96\sigma_Y) = 2.5\%$$

$$PR(Y > \mu_Y + 1.96\sigma_Y) = 2.5\%$$



→ $Z \sim N(0,1)$

$PR(Z \leq c) = \Phi(c)$

"PHI"

IF $Y \sim N(\mu_Y, \sigma_Y^2)$

THEN $\dfrac{Y - \mu_Y}{\sigma_Y} = Z \sim N(0,1)$

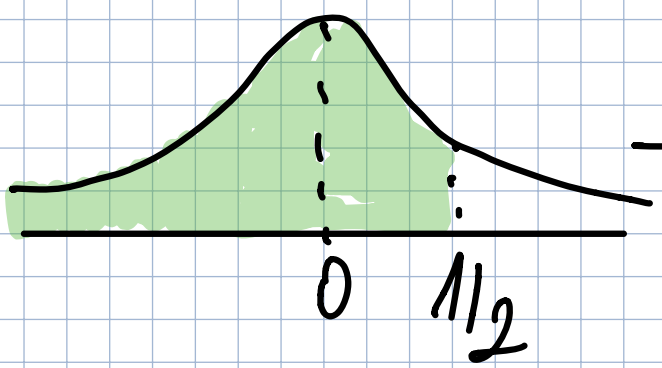
HANDON QUESTION: IF $Y \sim N(1,4)$

$$\frac{Y-1}{2} \sim N(0,1)$$

$$\text{so } Z \sim N(0,1)$$

$$PR(Y \leq 2) \rightarrow PR\left(\frac{Y-1}{2} < \frac{2-1}{2}\right)$$

$$\rightarrow PR(Z \leq 1/2)$$



$\rightarrow 0.691$

$$PR(1 \leq Y \leq 2) \rightarrow PR\left(\frac{1.1}{2} < \frac{Y-1}{2} < \frac{2-1}{2}\right)$$

$$PR(0 \leq Z \leq 1/2)$$

02/04/2021

- RECAP: LARGE SAMPLE APPROXIMATIONS TO SAMPLING APPROXIMATIONS
- LAW OF LARGE NUMBERS
- CENTRAL LIMIT THEOREM

SOLVE
$Y \sim N(1,4)$
$PR(1 \le Y \le 2)$

POPULATION ⟶ UNKNOWN BUT TRYING TO GATHER INFORMATION

↓ RANDOM SAMPLE n OBS [i.i.d]

(i) IDENTICALLY & (i) INDEPENDENTLY (D) DISTRIBUTED

$\{Y_1, Y_2, \ldots, Y_n\}$ ⟶ Y_i ⟶ EACH OBSERVATION

$$\bar{Y} = \sum_i Y_i / n = \frac{1}{n}(Y_1 + Y_2 + \ldots + Y_n)$$

LAW OF LARGE NUMBERS: AS n INCREASES

1. iid SAMPLE 2. FINITE VARIANCE

$n \to \infty$ (n>30) ⟶ STARTS TO WORK

$\bar{Y} \xrightarrow{P} E[Y]$

↓ APPROACH IN PROBABILITY

[WILL BE VERY CLOSE TO THE POPULATION MEAN]

CENTRAL LIMIT THEOREM: WITH LARGE ENOUGH n,

$$\bar{Y} \sim N\left(E[Y], \sigma_Y^2/n\right)$$

↳ THE INITIAL DISTRIBUTION OF Y DOES NOT MATTER

$Y \rightarrow$ BINARY (COIN)
UNIFORM (DIE)
NORMAL (HEIGHT)

$$\Big\} \bar{Y} \sim N\left(\mu_Y, \sigma_Y^2/n\right)$$

SIMULATION: $Y \sim N(5, 625)$

↳ σ_Y^2

$\downarrow$

$E[Y]$

1. $n = 100$

$$\bar{Y} \sim N\left(5, 625/100\right)$$

$$\sigma_{\bar{Y}}^2 = \sigma^2/100 = 625/100 = 6.25$$

$$\sigma_{\bar{Y}} = 2.5$$

2. $n = 500$ $\quad \sigma_{\bar{Y}}^2 = 625/500 = 1.25$

$$\sigma_{\bar{Y}} \cong 1.12$$

HANDOUT QUESTION: $Y \sim N(1,4)$

$PR(1 \leq Y \leq 2)$

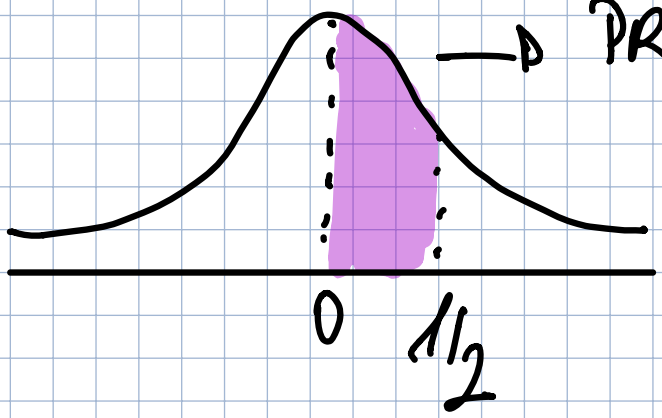
↳ RANDOM VARIABLE ITSELF.

↳ DICE ONE OBSERVATION FROM THE POPULATION.

$$PR(1 \leq Y \leq 2) \leftrightarrow PR\left[\frac{1-\mu_Y}{\sigma_Y} \leq \frac{Y-\mu_Y}{\sigma_Y} \leq \frac{2-\mu_Y}{\sigma_Y}\right]$$

↕ IDENTICAL

$PR(0 \leq Z \leq 1/2)$



$\rightarrow PR(Z \leq 1/2) - PR(Z \leq 0)$

$0.69146 - 0.5$

$= 0.19146$

$Y \sim ?(\mu_Y, \sigma_Y^2)$

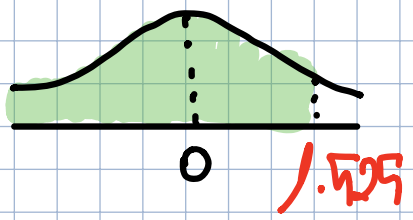
100 . 43

a. CLT

↳ $\bar{Y} \sim N(100.43/100)$

$$PR(\bar{Y} < 101) \rightarrow PR\left(\frac{\bar{Y}-\mu_Y}{\sigma_{\bar{Y}}} < \frac{101-100}{\sqrt{0.43}}\right.$$

$PR(Z < 1.525) = 0.9364$



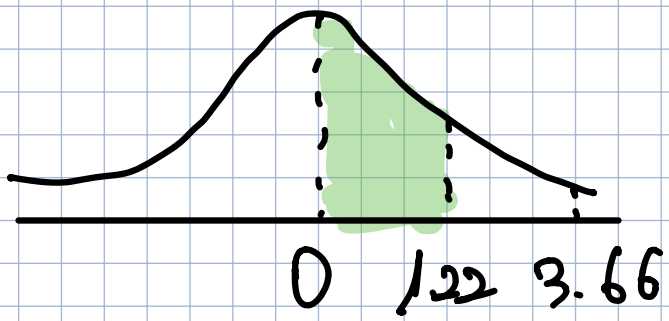
b. $n = 64$ $\quad PR(101 < \bar{Y} < 103)$

$$PR\left[\frac{101-100}{\sqrt{0.6719}} < \frac{\bar{Y}-100}{\sqrt{0.6719}} < \frac{103-100}{\sqrt{0.6719}}\right]$$

$\mu_Y = 100$

$\sigma_Y^2/n = 0.6719$

$\Rightarrow PR(1.22 < Z < 3.66]$



c. WHEN $n = 165$

$\sigma_{\bar{Y}}^2 = \sigma_Y^2/n = 113/165 = 0.2606$

$PR(\bar{Y} > 98) = 1 - PR(\bar{Y} < 98)$

$$1 - PR\left[\frac{\bar{Y}-100}{\sqrt{0.2606}} < \frac{98-100}{\sqrt{0.2606}}\right]$$

almost 100%

02/08/2021 CH 3 STATISTICS REVIEW

↳ PROPERTIES OF A GOOD ESTIMATOR

↳ HYPOTHESIS TESTING CONCERNING THE POPULATION MEAN

TRYING TO APPROXIMATE μ_Y , $E[Y]$ FOR RANDOM VARIABLE Y

$\hat{\mu}_Y$ → SAMPLE METRIC [CALCULATED USING AN i.i.d (RANDOM) SAMPLE]

WHAT IS A GOOD ESTIMATOR?

1. UNBIASED $E[\hat{\mu}_Y] = \mu_Y$
2. CONSISTENT $\hat{\mu}_Y \xrightarrow{P} \mu_Y$ AS $n \to \infty$
3. EFFICIENT (MINIMUM VARIANCE)

$VAR(\hat{\mu}_Y)$ SHOULD BE AS SMALL AS POSSIBLE.

$\bar{Y}$ → IS BLUE ESTIMATOR

↳ UNBIASED

↳ LINEAR

BEST

(MINIMUM VARIANCE — MOST PRECISE)

WHY? → CHOOSE AN m (ESTIMATOR) THAT MINIMIZE THIS

$\sum_i (Y_i - m)^2$

1. $\sum_i Y_i / n = \bar{Y}$ $\quad E[\bar{Y}] = E\left[\sum_i Y_i / n\right]$

$= 1/n \, E[\sum_i Y_i]$

$= 1/n \, E[Y_1 + Y_2 + \dots Y_n]$

$= 1/n \cdot n \mu_Y$ $\quad$ So $E[\bar{Y}] = \mu_Y$

$= \mu_Y$

2. $$\frac{d \sum_i (Y_i - m)^2}{dm} = -2 \sum_i (Y_i - m) = 0$$

$-2n(\bar{Y} - m) = 0$

$\bar{Y} = m$

HYPOTHESIS TESTING

TYPE I ERROR

TYPE II ERROR

SIGNIFICANCE LEVEL

CRITICAL VALUE

REJECTION ZONE

ACCEPTANCE ZONE

POWER OF THE TEST

P-VALUE

TWO SIDED VS ONE-SIDED ALTERNATIVES
CONFIDENCE INTERVALS

HANDOUT QUESTION

a. $Y = \{0, 1\}$ → FAIR COIN

$PR(Y=0) = 0.5$

$PR(Y=1) = 0.5$

b. $E[Y] = \sum_i Y_i \, PR(Y=Y_i)$

$= 0 \times PR(Y=0) + 1 \times PR(Y=1)$

$E[Y] = PR(Y=1) = 0.5$

$VAR[Y] = \sum_i (Y_i - \mu_Y)^2 \, PR(Y=Y_i)$

$= (0-0.5)^2 \times PR(Y=0) + (1-0.5)^2 \times PR(Y=1)$

$= 0.25 \times 0.5 + 0.25 \times 0.5$

$= 0.125 + 0.125 = 0.250$

$\sigma_Y^2 = 0.25 \quad E[Y] = \mu_Y$

$\sigma_Y = 0.5$ → STD. DEV

$Y \sim N(0.5, 0.25)$

c. CLT $\bar{Y} \sim N(\mu_Y, \sigma_Y^2/n)$

$n = 100$ $\bar{Y} \sim N(0.5, 0.25/100)$

$\sigma_{\bar{Y}}^2 = 0.0025$ $E[\bar{Y}] = 0.5$

$\sigma_{\bar{Y}} = 0.05$ → STD. DEV. OF SAMPLE MEAN

d. $n = 100$ 61 HEADS 39 TAILS ? TEST IF THE COIN IS FAIR

NULL HYP.

$H_0: E[Y] = \mu_{Y,0}$ → VALUE IS DETERMINED BY THE RESEARCHER

DEFAULT ASSUMPTION THAT YOU ARE TRYING TO REJECT

$H_A: E[Y] \neq \mu_{Y,0}$ → TWO-SIDED HYPOTHESIS

↓

ALTERNATIVE

$H_0: E[Y] = 0.5$ COIN IS FAIR

$H_A: E[Y] \neq 0.5$ COIN IS NOT FAIR

TWO POSSIBLE OUTCOMES: REJECT OR FAIL TO REJECT THE H_0

↳ NEVER 100% RELIABLE BECAUSE OF SAMPLE VARIATION

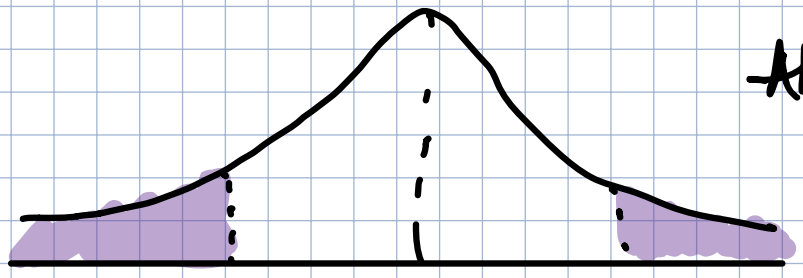
AFTER $n=100$ (PROVIDE A PROBABLE GUESS BASED ON SAMPLE)

HOW?

2. ASSUME NULL IS CORRECT AND USE CLT

IF $H_0: \mu_Y = 0.5$ THEN $\bar{Y} \sim N(0.5, 0.0025)$

AND $n=100$ → CALCULATE $\bar{Y}$ AND P-VALUE



$\mu_Y - (\bar{Y} - \mu_Y)$ μ_Y $\mu_Y + (\bar{Y} - \mu_Y)$

$$2 \times \Phi\left(-\left|\frac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}\right|\right) = \text{P-VALUE}$$

0.5

$H_0: \mu_Y = 0.5$

$H_1: \mu_Y \neq 0.5$

$\bar{Y} = 0.61$

P-VALUE = PROBABILITY TO DRAW A SAMPLE WITH $\bar{Y}$ MEAN THAT IS $|\mu_{Y,0} - \bar{Y}|$ ABOVE OR BELOW μ_Y IF H_0 IS TRUE

P. VAL : THE LIKELIHOOD OF DRAWING A SAMPLE $n=100$ THAT HAS A SAMPLE MEAN THAT IS 0.11 ABOVE | BELOW

IF H_0 IS CORRECT

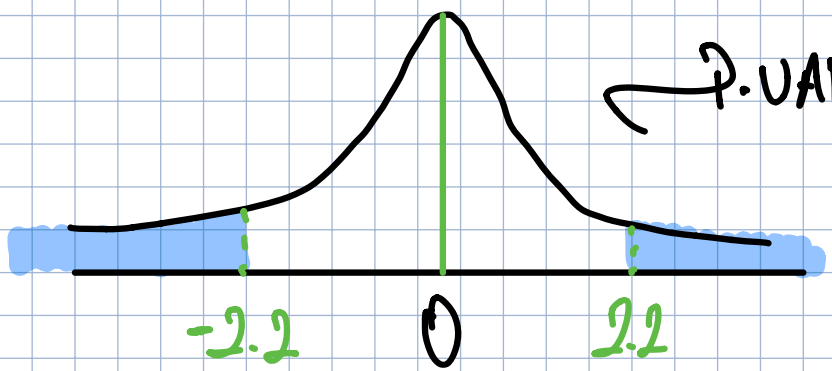
P. VAL CALCULATION

1. • σ_Y^2 IS KNOWN
2. • σ_Y^2 IS UNKNOWN

1. → Z-SCORE $= \dfrac{\bar{Y} - \mu_{Y,0}}{\sigma_{\bar{Y}}}$ $\overset{\text{IF } H_0}{\sim}$?

$= \dfrac{0.61 - 0.50}{0.05} = \dfrac{0.11}{0.05} = 2.2$

↳ WE KNOW THIS BC $Y \sim B(0.5, 0.25)$



↳ P. VAL $= 2 \times \Phi[-|2\text{-SCORE}|]$

↳ STD. NORMAL DIS.

$2 \times \Phi(-2.2)$

$= 2 \times 0.0139$

$= 0.0278$

IF THE COIN IS FAIR, THEN THERE IS A 2.78% CHANCE THAT WITH $n=100$ $\bar{Y} > 0.61$ & $\bar{Y} < 0.39$

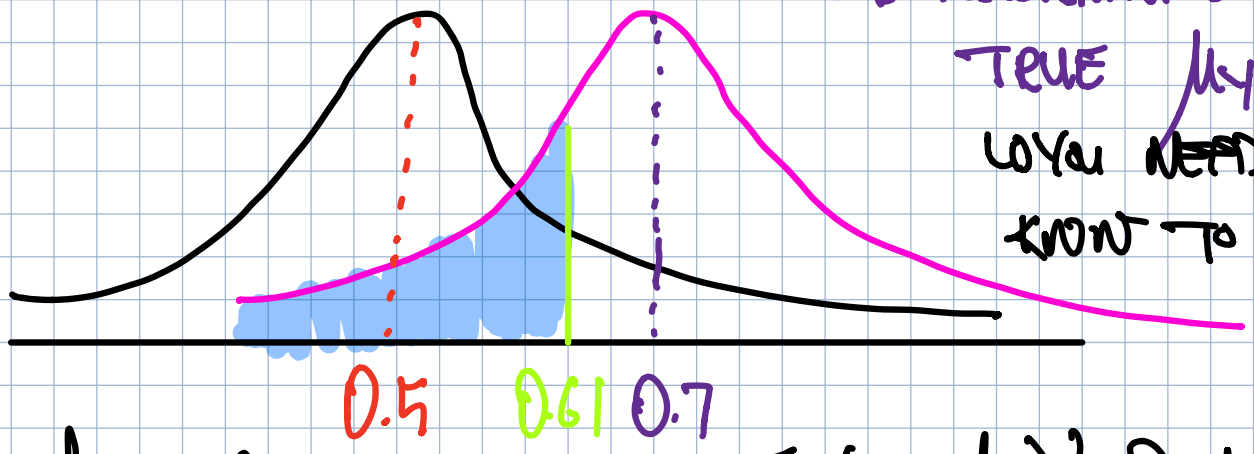
f. P-VAL: TYPE-I ERROR → α (ALPHA LEVEL)

↳ H_0 IS CORRECT

↳ BUT YOU REJECT IT.

g. TYPE-II ERROR: H_0 IS FALSE BUT YOU FAIL TO REJECT IT

FALSE NEGATIVE



→ ALTERNATIVE H_A TRUE $\mu_Y = 0.7$

↳ YOU NEED TO KNOW TO CALCULATE TYPE II.

$E[\mu_Y] = 0.70$ $\quad VAR(Y) = \sum_i (Y_i - \mu_Y)^2 \cdot PR(Y = Y_i)$

$\sigma_Y^2 = (0-0.7)^2 \times 0.30 +$

$(1-0.7)^2 \times 0.70$

$\sigma_Y^2 = 0.21$

$\sigma_{\bar{Y}} = 0.21/100$

$= 0.0021$

$\bar{Y} \sim N(0.70, 0.0021)$

$\Phi\left[-\left|\frac{0.61-0.70}{\sqrt{0.0021}}\right|\right]$

$\Phi(-1.96) \cong 2.5\%$

h. IF σ_Y^2 IS UNKNOWN, USE s_Y^2 AS AN ESTIMATOR FOR σ_Y^2

$$s_Y^2 = \frac{1}{n-1}\sum_i (Y_i - \bar{Y})^2 \rightarrow \text{SAMPLE VARIANCE}$$

s_Y = SAMPLE STD. DEV $= \sqrt{s_Y^2}$

$$\sigma_{\bar{Y}}^2 = \frac{\sigma_Y^2}{n} \leftrightarrow \sigma_{\bar{Y}} = \frac{\sigma_Y}{\sqrt{n}}$$

$$SE[\bar{Y}] = \hat{\sigma}_{\bar{Y}} = s_Y/\sqrt{n}$$

↳ STANDARD ERROR OF $\bar{Y}$

$$t\text{-stat} = \frac{\bar{Y} - \mu_{Y,0}}{SE[\bar{Y}]}$$

$s_Y = 0.5$

$\hat{\sigma}_{\bar{Y}} = 0.049$

$$t\text{-stat} = \frac{0.61 - 0.5}{0.049} \simeq 2.244$$

↳ P-VALUE $= 0.02706281$

WITHIN THIS INTERVAL

4. 95% CI $\rightarrow \bar{Y} \pm 1.96\, SE[\bar{Y}]$

$0.61 \pm 1.96 \times 0.049$ $\quad [0.51 \quad 0.70]$

95% PROBABILITY THAT THE TRUE PARAMETER μ_Y

POWER OF A TEST = 1. PR (TYPE II ~~ERROR~~)

↳ PROBABILITY THAT TEST CORRECTLY REJECTS THE H_0

H_0 IS FALSE BUT YOU FAIL TO REJECT IT

COMPARING MEANS FROM TWO DIFFERENT POPULATIONS:

X & Y → TWO ~~DIFFERENT~~ POPULATIONS (RANDOM VARIABLES)

μ_Y μ_X

$\bar{Y}$ $\bar{X}$

S_Y^2 S_X^2

$SE[\bar{Y}]$ $SE[\bar{X}]$

$\bar{X} \sim N(\mu_X, \sigma_X^2/n_X)$

$\bar{Y} \sim N(\mu_Y, \sigma_Y^2/n_Y)$

IF X & Y ARE BOTH i.i.d

$(\bar{X}-\bar{Y}) \sim N\left(\mu_X-\mu_Y, \frac{\sigma_X^2}{n_X}+\frac{\sigma_Y^2}{n_Y}\right)$

UNKNOWN

$$SE[\bar{X}-\bar{Y}] = \sqrt{\frac{S_X^2}{n_X}+\frac{S_Y^2}{n_Y}}$$

$H_0: \mu_X - \mu_Y = d_0$

$H_A: \mu_X - \mu_Y \neq d_0$

$$t = \frac{(\bar{x} - \bar{y}) - d_0}{SE[\bar{x} - \bar{y}]}$$

↳ P-VAL = 2 x Φ (- |t|)

CONFIDENCE INTERVAL :

$$(\bar{y} - \bar{y}) \mp 1.96 \times SE[\bar{x} - \bar{y}]$$

QUESTION 2:

$\mu_{Y,0} = 15.4$ → HYP0

$\bar{Y} = 14.6$

$\sigma_Y = 2.5$

$n = 35$

$$\sigma_{\bar{Y}} = 2.5 / \sqrt{35}$$

$$z = (\bar{Y} - \mu_{Y,0}) / \sigma_{\bar{Y}}$$

02/18/2021 LESSON PLAN:

- COMPARING THE MEANS FROM 2 POP.
- SOLVE HANDOUT QUESTION
- ANSCOMBE'S QUARTET
- INTRODUCTION TO REGRESSION ANALYSIS (CHAPTER 4)

COMPARING SAMPLE MEANS FROM 2 DIFFERENT POPULATIONS

$H_0: \mu_X - \mu_Y = d_0$

$H_A: \mu_X - \mu_Y \neq d_0$

X & Y — RANDOM VAR DIFF. POP.

$\bar{X}, \bar{Y}$

σ_X^2, σ_Y^2

SE[X̄] SE[Ȳ]

CLT:

$\bar{X} \sim N(\mu_X, \sigma_X^2/n_X)$ ← SAMPLE SIZE FOR X

$\bar{Y} \sim N(\mu_Y, \sigma_Y^2/n_Y)$ ← SAMPLE SIZE FOR Y

$$(\bar{X} - \bar{Y}) \sim N\left(\mu_X - \mu_Y, \frac{\sigma_X^2}{n_X} + \frac{\sigma_Y^2}{n_Y}\right)$$

TYPICAL PROBLEM: σ_X^2, σ_Y^2 NOT OBSERVED

ASSUMPTION: X & Y ARE BOTH iid

$$SE[\bar{x}-\bar{y}] = \sqrt{\frac{\sigma_x^2}{n_x} + \frac{\sigma_y^2}{n_y}}$$

$$t = \frac{(\bar{x}-\bar{y}) - d_0}{SE[\bar{x}-\bar{y}]}$$

$$P\text{-val} = 2 \times \Phi(-|t\text{-stat}|)$$

CONFIDENCE INTERVALS $\} \; (\bar{x}-\bar{y}) \pm 1.96 \times SE[\bar{x}-\bar{y}]$

s_x^2 → SAMPLE VARIANCE $= \sum_i (x_i-\bar{x})^2 / (n_x-1)$

s_y^2 → SAMPLE VARIANCE $= \sum_i (y_i-\bar{y})^2 / (n-1)$

s_{xy} → SAMPLE COVAR $= \dfrac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{(n-1)}$

↓

σ_{xy} → POPULATION COVARIANCE

→ SAMPLE COR.

$$\rho_{xy} = \quad -1 \le \rho_{xy} \le 1$$

CHAPTER 4: REGRESSION ANALYSIS

$$Y_i = \beta_0 + \beta_1 X_i + \mu_i$$

RESPONSE VAR
LEFT-HANDSIDE VARIABLE

$i = \{1, \ldots, n\}$ → SAMPLE SIZE

Y = DEPENDENT VARIABLE
OUTCOME VARIABLE

X → INDEPENDENT VARIABLE
RIGHT HANDSIDE VARIABLE
REGRESSOR

β_0, β_1 → POPULATION COEFFICIENTS
PARAMETERS

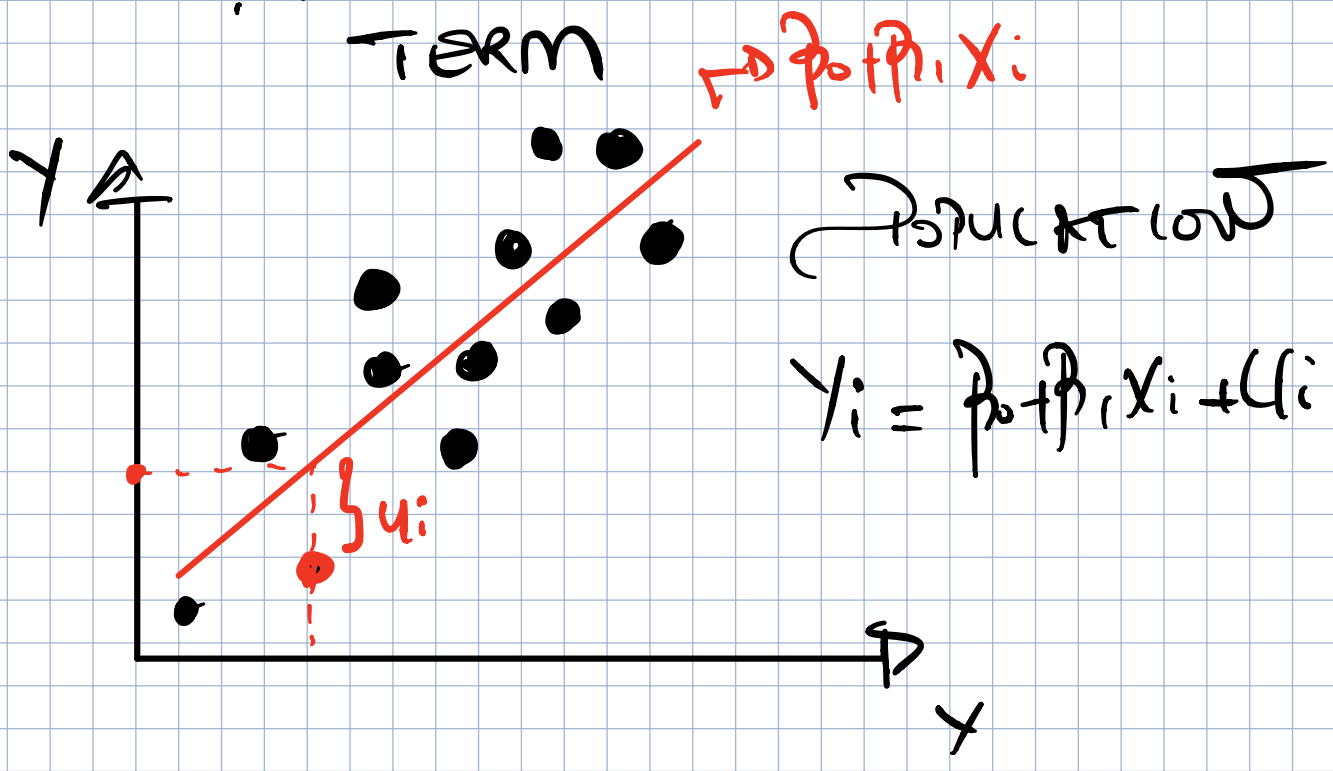
β_0 : INTERCEPT

β_1 : SLOPE

$\beta_0 + \beta_1 X_i$ → POPULATION REGRESSION LINE

POPULATION REGRESSION FUNCTION

u_i :→ ERROR TERM



POPULATION

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

BUT POPULATION PARAMETERS β_0, β_1 CAN'T BE KNOWN BECAUSE WE DON'T OBSERVE THE POPULATION DATA.

TWO MAIN PROBLEMS:

(RETHINK ABOUT IT LATER)

1° POPULATION PARAMETERS → SAMPLE

2° HOW TO FIT THE BEST LINE?

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

↓ → BIRTH WEIGHT

EARNINGS

CHILD'S SCHOOL ACHI. → SALES, ...

→ PARENTS INCOME

→ SPENDING ON SOCIAL MEDIA AD

USING A SAMPLE OF n OBSERVATIONS.

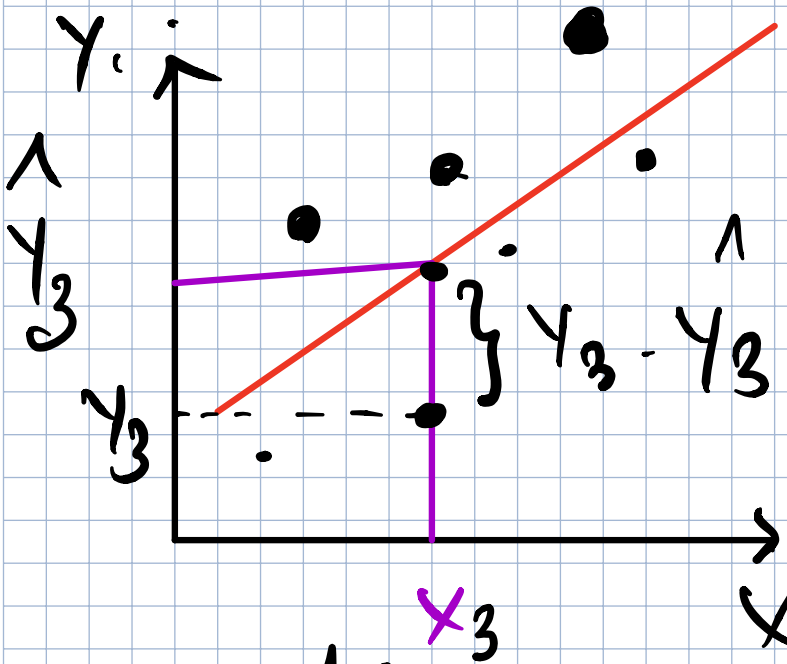
CHOOSE $b_0 \to \beta_0$

$b_1 \to \beta_1$ [2 ESTIMATORS]

THAT BEST DESCRIBE THE RELATIONSHIP BTW X_i & Y_i?

i	X	Y	$\hat{Y}$	$\hat{u}$
1	x_1	y_1	$\hat{y}_1$	$\hat{u}_1$
2	x_2	y_2	$\hat{y}_2$	$\hat{u}_2$
⋮	⋮	⋮	⋮	⋮
n	x_n	y_n	$\hat{y}_n$	$\hat{u}_n$

THE LINE THAT MAKES THE LEAST POSSIBLE AMOUNT OF PREDICTION MISTAKE.



→ PREDICTION MISTAKE

$$\sum_{i=1}^{n} \hat{u}_i^2 = \sum_i (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$$

CHOOSE $\hat{\beta}_0, \hat{\beta}_1$ TO MINIMIZE THE SUM OF SQUARED RESIDUALS

ORDINARY LEAST SQUARES

$$\hat{\beta}_1 = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_i (X_i - \bar{X})^2} = \frac{S_{XY}}{S^2_X}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad i = 1, \ldots, n$$

$$\hat{u}_i = Y_i - \hat{Y}_i \qquad i = 1, \ldots, n$$

STANDARD ERROR OF THE REGRESSION

IS MEASURE OF THE SPREAD OF THE OBSERVATIONS AROUND THE REGRESSION LINE.

$$SER = S_{\hat{u}} = \sqrt{\sum \hat{u}_i^2 \Big/ (n-2)} = \sqrt{\frac{SSR}{(n-2)}}$$

MAGNITUDE OF A TYPICAL PREDICTION ERROR

IN $\hat{\beta}_0, \hat{\beta}_1$ SHOULD BE KNOWN BEFORE SO LOSE 2 D.F

MEASURES OF FIT: HOW GOOD THE REGRESSION LINE FITS THE DATA?

R^2 → R-SQUARED MOST COMMON

TSS → TOTAL SUM OF SQUARES

$TSS = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$ → TOTAL VARIATION IN Y

SSR (RSS) = SUM OF SQUARED RESIDUALS
RESIDUAL SUM OF SQUARES

$\sum \hat{u}_i^2$

$ESS = TSS - SSR = \sum_i (\hat{Y}_i - \bar{Y})^2$

UNEXPLAINED SUM OF SQUARES

$Y_i = \hat{Y}_i + \hat{u}_i$

$R^2 = \frac{ESS}{TSS}$ → THE VARIATION IN Y EXPLAINED BY THE VARIATION IN X

$0 \leq R^2 \leq 1$

SOME FUN FACTS ABOUT OLS

1. $1/n \sum_i \hat{u}_i = 0$
2. $1/n \sum_i \hat{y}_i = \bar{y}$ — REGRESSION LINE PASSES THROUGH $\bar{x}, \bar{y}$
3. $\sum_i \hat{u}_i x_i = 0$
4. $\sum_i \hat{u}_i^2$ MINIMUM
5. $TSS = ESS + SSR$

STANDARD ERROR OF THE REGRESSION

IS MEASURE OF THE SPREAD OF THE OBSERVATIONS AROUND THE REGRESSION LINE.

$$SER = S_{\hat{u}} = \sqrt{\sum \hat{u}_i^2 / (n-2)} = \sqrt{\frac{SSR}{(n-2)}}$$

MAGNITUDE OF A TYPICAL PREDICTION ERROR

($\hat{\beta}_0, \hat{\beta}_1$ SHOULD BE KNOWN BEFORE SO LOSE 2 D.F)

LECTURE PLAN 03/08/22

i. OLS ASSUMPTIONS (SPECIFIC FOCUS ON THE CAUSAL INTERPRETATIONS OF OLS COEFFICIENTS $\hat{\beta}_0, \hat{\beta}_1$

ii. SAMPLING DISTRIBUTION OF THE OLS ESTIMATOR.

iii. GO OVER A FEW PTS ON THE HOMEWORK ASSIGNMENTS

↳ PROBABLY ON NEXT CLASS

iv: CONFIDENCE INTERVALS & HYPOTHESES TEST USING (SINGLE REGRESSION)

$$Y_i = \beta_0 + \beta_1 X_i + u_i$$

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$$

$$Y_i = \hat{Y}_i + \hat{u}_i$$

→ RANDOM SAMPLE of n OBSERVATIONS

→ SAMPLING VARIATION ($Y_i, \hat{Y}_i, \hat{u}_i$)

THE LEAST SQUARES ASSUMPTIONS

1. OLS COEFFICIENTS $\hat{\beta}_0$ & $\hat{\beta}_1$ HAVE NO CAUSAL INTERPRETATION UNLESS $E[u_i | X_i] = 0$

$E[u_i, x_i] = \text{cor}(u_i, x_i) = 0$

↳ ASSUMPTION IS ABOUT THE POPULATION SO NO PERFECT WAY TO TEST

$Y_i = \beta_0 + \beta_1 X_i + u_i$

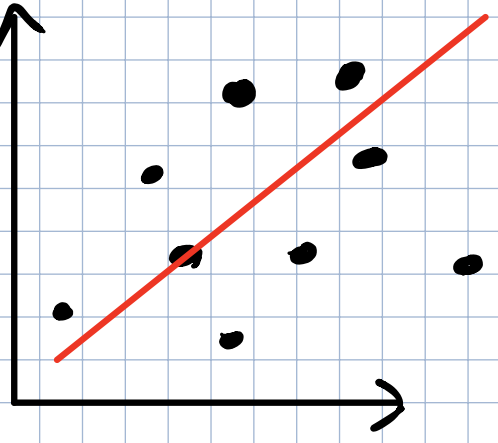
X_i & Y_i ARE ASSOCIATED BUT IT DOES NOT MEAN THAT $X_i \rightarrow Y_i$)

CAUSES

$Y_i = \beta_0 + \beta_1 X_i + u_i$

Y_i ↓ HEALTH (LIFE EXPECTANCY)

X_i ↓ EDUCATION (YEARS OF SCHOOLING)



$\beta_1 = 0.15$ ↑ AN ADDITIONAL OF EDUCATION IS ASSOCIATED WITH A 0.15 YEARS OF LIFEEXPECTANCY.

↳ WHEN CONCLUDE CAUSE)

2 ISSUES:

1. REVERSE CAUSALITY Y CAUSING X

HEALTH → EDUCATION

2. THIRD OMITTED VARIABLE THAT IMPACT BOTH

↳ EDUCATION

↳ HEALTH

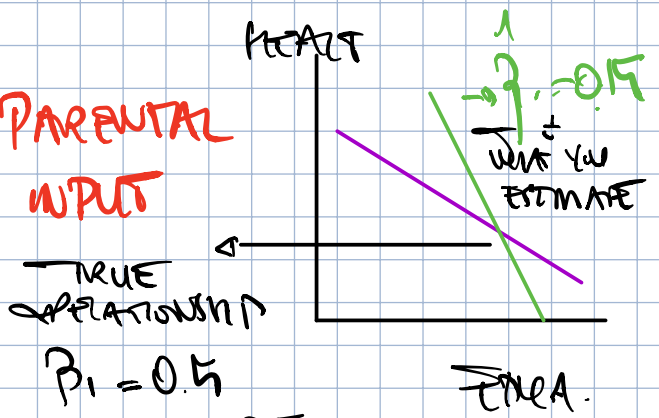
↳ PARENTAL INPUT

$Y_i = \beta_0 + \beta_1 X_i + u_i$

↓ HEALTH

↓ EDUCATION

↳ PARENTAL INPUT



IF $COR(X_i, u_i) = 0$ → β_1 CAN ONLY BE DUE TO

↳ N. OMITTED VARIABLE BIAS → IF $COR(X_i, u_i) \neq 0$

X_i $E[u_i | X_i] = 0$ (IDENTICAL)

LISTEN NO CAUSAL INTERPRETATION

2°

OLS ASSUMPTION 2° X_i, Y_i $i = 1, \ldots, n$

ARE i.i.D → DISTRIBUTED ↓ IDENTICALLY INDEPENDENTLY

} RANDOM SAMPLING INSURES THIS ASSUMPTION

3° LARGE OUTLIERS ARE UNLIKELY

(IF NOT OLS MIGHT BE MISLEADING)