CHAPTER 1:

* Def: Statistics:
  - Collecting
  - organizing
  - analyzing
  - interpreting

  } => make conclusion

* purpose: change "Data" into conclusion

* Classification:

Statistics { Descriptive: organizing, summerizing and display
Differential: using sample data to drain conclusion about to population

* Voc
- population (N): everything, everyone
- parameter: characteristics of population
- saple sample(n): a portion of population
- statistics : description of sample
- variable : characteristics of elements
- data : actual value of variables

* Exp:

A	$ 1000
B	$ 1500
C	$ 2000

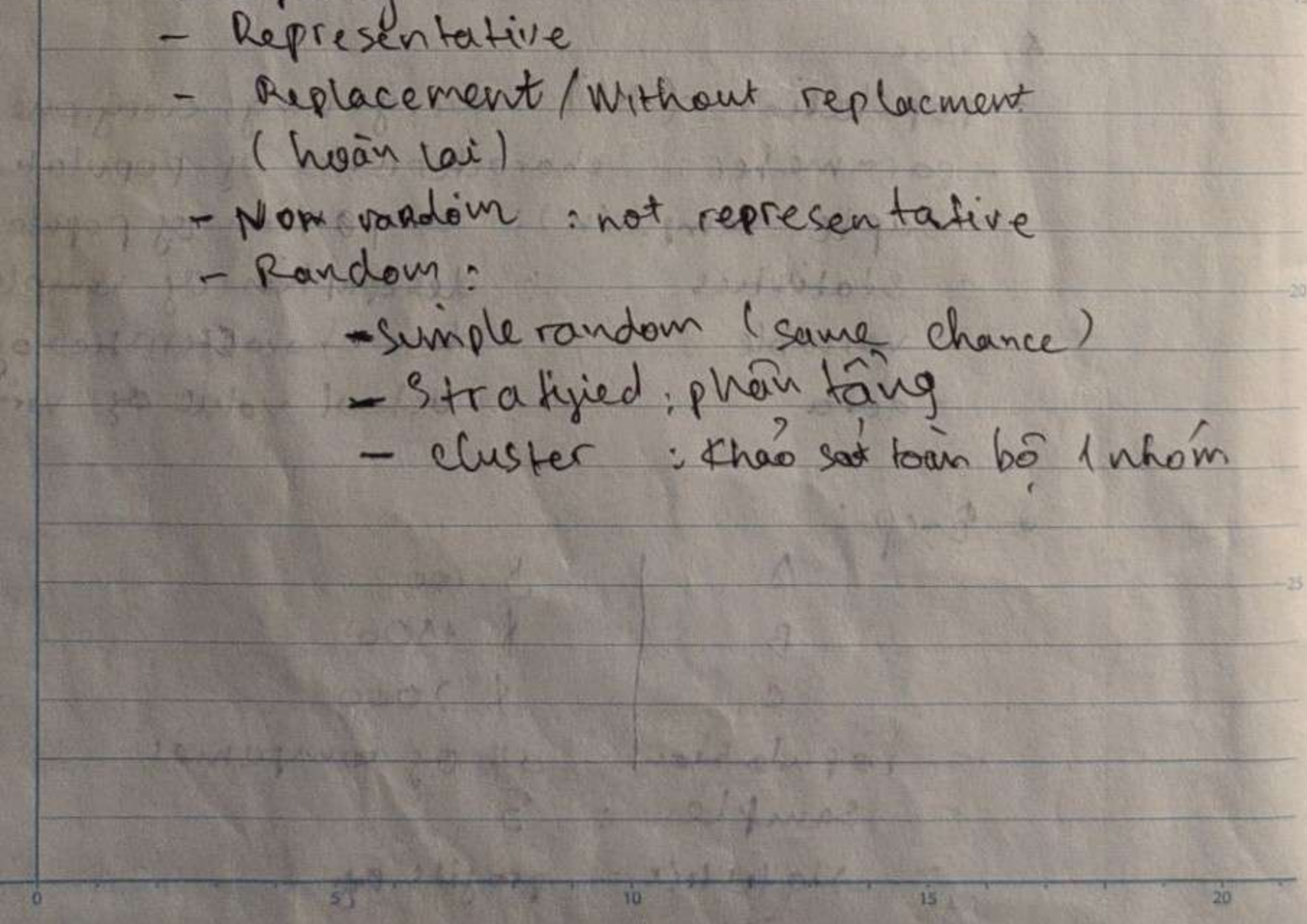
- Population: all of companies
- sample : 3
- Statistic : profits of
- variable : profits of each company

Thứ
Ngày . . No.

* collect data
  1, Retrospective (Historical): lấy data ở quá khứ
  2, Observational Study
  3, Experiment
  4, Simulation
  5, Survey [sample : 50
              population : CENSUS

* Type of data
  - Qualitative :
  - Quantitative :
    + Counting : # Students, # cars, ...
    + measure : temperature, weight, height, salar

* Sampling
  - Representative
  - Replacement / Without replacment
    (hoàn lại)
  - Non random : not representative
  - Random :
    - simple random (same chance)
    - Stratified : phân tầng
    - cluster : khảo sát toàn bộ 1 nhóm



CHAPTER 2: Probability

Ex: Random experiment

toss a coin : - outcomes [head, tail

- sample space : all of outcomes = $\Omega = \{H, T\}$

- event : + simple : $B_1 = \{H\}$
  + compound : $B_2 = \{HH\}$

(+) event is a subset of the sample ~~sap~~ space of a random / is a collection of outcomes / is a collection of elementary of events

Ex: Roll a dice

$\Omega = \{1, 2, 3, 4, 5, 6\}$

$A = \{1, 3, 5\}$

* $P(A) = \frac{|A|}{|\Omega|}$

* Properties :

$0 \leq P(A) \leq 1$

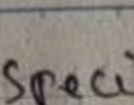
impossible → 0 — 0,05 — 0,5 — 1 * certain

unusual ↗ 50-50

* Some probability formulas

① Additional (OR)

$P(A \cup B) = P(A) + P(B) - P(A \cap B)$

- Special: 

$P(A \cup B) = P(A) + P(B)$

② Multiplication (AND)

- If A and B are independent $\Rightarrow P(A \cap B) = P(A) \cdot P(B)$

- General case:

$$P(A \cap B) = P(A)\,P(B|A)$$
$$= P(B)\,P(A|B)$$

③ $P(A|B) = P(A)$ given that B

④ Complement:

$$P(\bar{A}) = 1 - P(A)$$

Ex:	smoke	not smoke
men	200	150
women	85	115

a) $P(\text{men}) = \frac{7}{11}$ $\frac{57}{110}$

b) $P(\text{smoke}) =$ $\frac{200}{285}$

c) $P(\text{men} | \text{smoke}) =$ $\frac{200}{550} = \frac{4}{11}$

d) $P(\text{men and smoke}) =$ $\frac{150}{265}$

e) $P(\text{men} | \text{not smoke}) =$

Ex:

	under 21	21-25	over 25		
ticket	82	39	18	13	9
no ticket	17	27	61	10	5
	99	66	79	24	9

a)

b) $P(\text{no ticket}) = \frac{105}{209}$

c) $P(21\text{-}25 \text{ and no ticket}) = \frac{22}{209}$

d) $P(21\text{-}25 \text{ or no ticket}) = \frac{199}{299}$

e) $P(21\text{-}25 \mid \text{no ticket}) = \frac{27}{105}$

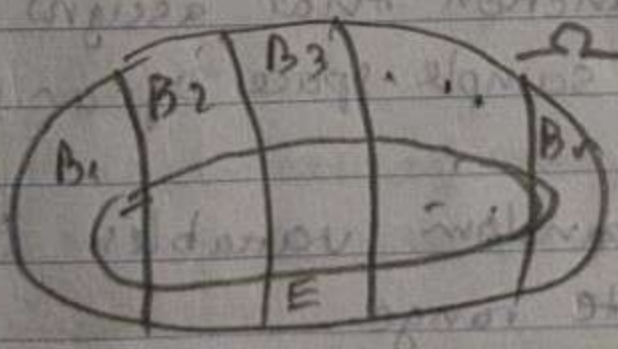
f) $P(\text{no ticket} \mid 21\text{-}25) = \frac{27}{66}$

Note:

A, B independent

$\rightarrow \begin{cases} P(A \cap B) = P(A)\,P(B) \\ P(A) = P(A \mid B) \\ P(B) = P(B \mid A) \end{cases}$

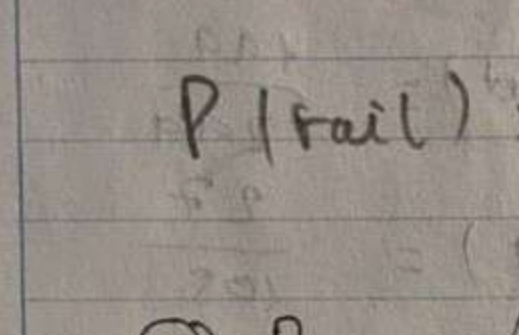
* Total probability



$$P(E) = P(E \mid B_1)\,P(B_1) + P(E \mid B_2)\,P(B_2) + \dots + P(E \mid B_n)\,P(B_n)$$

Thứ
Ngày . . No.

Ex: Build tree



$P(fail) = 90\%.5\% + 10\%.20\%$
$= 6.5\%$

⑧ Bayes's
$P(A|B) = P(B|A).P(A) / P(B)$

CHAPTER 3:
Discrete random variables

* Random variables
  - describle the outcome x of an experiment
  - the function that assigns a number to each out come in the sample space of an random experiment

⑧ Discrete random variables (DRV)
  - a injinite range

⑧ Probobility distribution junction (PDF)

X	0	1	2	3	4	5
P(x)	0.2	0.1	X	0.3	0.1	0.1

Thứ
Ngày . . No.

⊛ Mean and variance (kì vọng và phương sai)

$$\mu = E(x) = \sum_{1}^{N} x_i P(X = x_i) = \Sigma \frac{x_i}{N}$$

$$\sigma^2 = V(x) = E(x-\mu) = \sum_{i=1}^{N} (x_i - \mu)^2 P(x = x_i)$$

$$= \sum_{i=1}^{N} x_i^2 P(X = x_i) - \mu^2$$

$$\sigma = \sqrt{\sigma^2} = \sqrt{V(x)}$$

Ex: 5, 6, 7, 8

$$\sigma = \sqrt{5^2 \cdot \frac{1}{4} + 6^2 \cdot \frac{1}{4} + 7^2 \cdot \frac{1}{4} + 8^2 \cdot \frac{1}{4}} - 6,5$$

⊛ Empirial Rule

(68 - 95 - 99,78)

Ex: The number of messages sent per hour over a computer network has the following distribution

X	9	10	11	12	13	14
P(X)	0.08	0.15	0.3	X	0.2	0.07

a) Find avg / mean / expected value
b) variance
c) standard deviation

a) $\mu = 1,5$

b) $\sigma^2 = 1.85$

c) $\sigma = 1.36$

* Probability mass function

$f(x_i) = P(x = x_i)$, $0 \leq P(x = x_i) \leq 1$

$\Downarrow$

$\begin{cases} 0 \leq f(x_i) \leq 1 \\ \Sigma f(x_i) \end{cases}$

Exc :

x	0	1	2
$P(x)$	25/36	10/36	1/36

$f(0) = P(X=0) = 25/36$

$f(1) = P(X=2) = 10/36$

$f(2) = P(X=3)$

- Commulative distribution

$F(a) = P(x \leq a) = \sum_{x_i \leq a} P(x = x_i)$

$= \sum_{x_i \leq a} f(x_i)$

$\Downarrow$

$\begin{cases} 0 \leq f(x) \leq 1 \\ F: \text{not continuous} \end{cases}$

Thứ
Ngày . . No.

Ex 1: $F(0) = P(x \leq 0) = P(x = 0) = \frac{25}{36}$

$F(1) = P(x < 1) = P(x = 1) - P(x = 0)$

$- P(a \leq x \leq b) = P(x \leq b) - P(x < a)$

Ex 2: Determine the probability mass function of x from the following cummulative distribution function ($F(x) \to f(x)$)

$$F(x) = \begin{cases} 0, & x < -2 \\ 0.2, & -2 \leq x < 0 \\ 0.7, & 0 \leq x < 2 \\ 1, & x \geq 2 \end{cases}$$

a) find $P(0 \leq x \leq 2)$
b) find $P(x < 1.7)$

$f(-2) = 0.2 - 0 = 0.2$
$f(0) = 0.7 - 0.2 = 0.5$
$f(2) = 1 - 0.7 = 0.3$

KOKUYO

Thứ

Ngày . . No.

* Review: - Mass function

$f(x_i) = P(x = x_i)$

→ $\begin{cases} 0 \le f(x_i) \le 1 \\ \sum_{i}^{N} f(x_i) = 1 \end{cases}$

- Cumulative destribution function

$$F(a) = P(X \le a) = \sum_{x \le a} P(x = x_i) = \sum_{x \le a} f(x_i)$$

- $E(ax + by) = aE(X) + b(Ey)$
- $V(ax+by) = a^2 V(x) + b^2 V(y)$

(!) $V(ax+b) = a^2 V(x)$

* Special distribution

① Discrete Uniform Distribution

$X = \{x_1, x_2, \dots, x_n\}$

$f(x_i) = 1/n = P(X = x_i)$

KOKUYO

Ex: Rolling a dice $\rightarrow x \in \{1,2,3,4,5,6\}$

$$E(x) = \sum x_i P(x = x_i) = 1 \cdot \frac{1}{6} + \frac{2}{1} \cdot \frac{1}{6} + \dots + 6 \cdot \frac{1}{6} = \mu$$

$$V(X) = \sum x_i^2 P(x = x_i) - \mu^2$$
$$= 1^2 \cdot \frac{1}{6} + 2^2 \cdot \frac{1}{6} + \dots + 6^2 \cdot \frac{1}{6} - \mu^2$$

1.2 ⊗ $\begin{cases} E(X) = \frac{a+b}{2} \\ V(X) = \frac{(b-a+1)^2 - 1}{2} \end{cases}$ (chỉ dùng trong tập rời rạc liên tiếp)

X: discrete uniform random variable on the consecutive integer $[a, b]$

Ex 1: X: uniform discrete random over $[10, 60]$

σ (Standard deviation) of X?

$$\sigma = \sqrt{V(X)} = \sqrt{E(X-\mu)^2} = \sqrt{\frac{(60-10+1)^2 - 1}{12}}$$
$$= \frac{5\sqrt{78}}{3}$$

Ex 2:

$X \sim U[2,5]$. $V(4x)$?

$$V(4x) = 16\, V(x) = 16 \cdot \frac{(5-2+1)^2 - 1}{12}$$

$$= 20$$

Ex 3: X {1, 2, 3, 6, N} equally

Mean = 10

find N?

$$\frac{1+2+3+6+u}{5} = 10$$

$$\Rightarrow u = 38$$

② Binomial Distribution ($X \sim B(n,p)$)
(What is the probability of k success out of n trial)
Ex: 9 student, P(success) = 0.3
P(3 students: success) = ? in n
=> X: number of k success ~~student~~ trial,
p: prob of success, q: prob of fail

$$P(X = k) = C_n^k \cdot p^k \cdot q^{n-k} \quad (1-p)^{n-k}$$

* Mean and variance ($X \sim B(n,p)$)

$E(X) = np$

$V(X) = npq$

$P(X > k) = 1 - p(X \leq k)$

$P(X \geq k) = 1 - P(X \leq k-1)$

Ex: dice rolled 20 times, X be the numbers of ⚅ come up. Find the σ

$$\sigma = \sqrt{V(X)} = \sqrt{npq} = \sqrt{20 \cdot \frac{1}{6} \cdot \frac{5}{6}}$$

③ Poisson
(what is the prob of n of success in a fixed interval: time, area, ...)

- X = number of events that can appear in a fixed interval with parameter $\lambda > 0$
- λ = The average of events / interval
(mean) (per)

EXCEL: $P(X = k) =$ POISSON.DIST(k, λ, T)

$$\Rightarrow P(X=k) = \frac{e^{-\lambda T}}{k!} \cdot (\lambda T)^k \quad k = 0, 1, \dots$$

3.1 Mean and Variance

$$E(X) = \lambda T$$
$$V(X) = \lambda T$$

④ Hypergeometric distribution

(what is the probability of the number of success in sample

N

(k success, N-k fail) —sample n, without replacement→ X = num of success in n sample

$X = \{1, 2, 3, \dots, n\}$

$$(l \le n) \quad P(X=l) = \frac{C_K^l \, C_{N-K}^{n-l}}{C_N^n}$$

$$E(X) = np \quad \left(p = \frac{K}{N}\right)$$

$$V(X) = npq \frac{N-n}{N-1}$$

Ex: 30 red, 20 green, without replacement choose 5 balls, $E(X)$ and σ at the number of red balls

$$E(X) = np = 5 \cdot \frac{30}{50} = 3$$

$$\sigma = \sqrt{V(X)} = \sqrt{5 \cdot \frac{3}{5} \cdot \frac{2}{5} \cdot \frac{50-5}{50-1}} \approx 1.05$$

Ex 2: $N = 36$, fail: 12,
$n = 7$
$P(3)$?

$$P(3) = \frac{C_{29}^{3} \cdot C_{12}^{7-3}}{C_{N=36}^{n=7}}$$

(5) Geometric Distribution
(what is the probability of the number of trials until the 1st success before)

X = number of trails until the first success

$X \sim$ geometric $\rightarrow P(X=k) = (1-p)^{k-1} \cdot p$

$$E(X) = \frac{1}{p}$$

$$V(X) = \frac{1-p}{p \times p}$$

Ex: $P(\text{error}) = 0.03$

a) $P(X=5) = (1-0.03)^{5} \cdot 0.03$
$= 0.0257$

b) $P(X \leq 5) = P(X=1) + P(X=2) + P(X=3) + P(X=4) + P(X=5)$
$= (0+p) + (1-p)p + (1-p)^2 p + (1-p)^3 p + (1-p)^4 p$

X C V B N M < > ?

Thứ
Ngày No

(6) Negative Binomial distribution
(What is the probability of the number of trial until r^{th} success)
X = number of trials until r^{th} success
~~X = { r, r+1, ..., ∞ }~~

$$P(X = k) = C_{k-1}^{k-1} p^r q^{k-r}$$

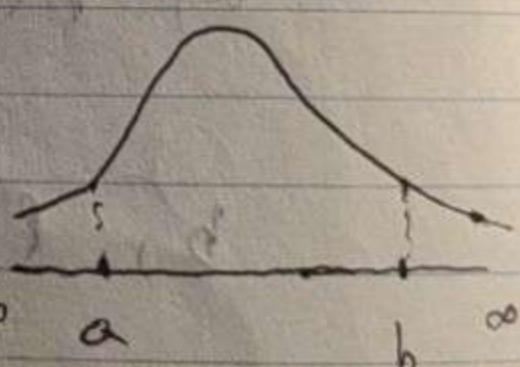
* Mean and variance
$$E(x) = \frac{r}{p}$$
$$V(x) = \frac{r(1-p)}{p^2}$$

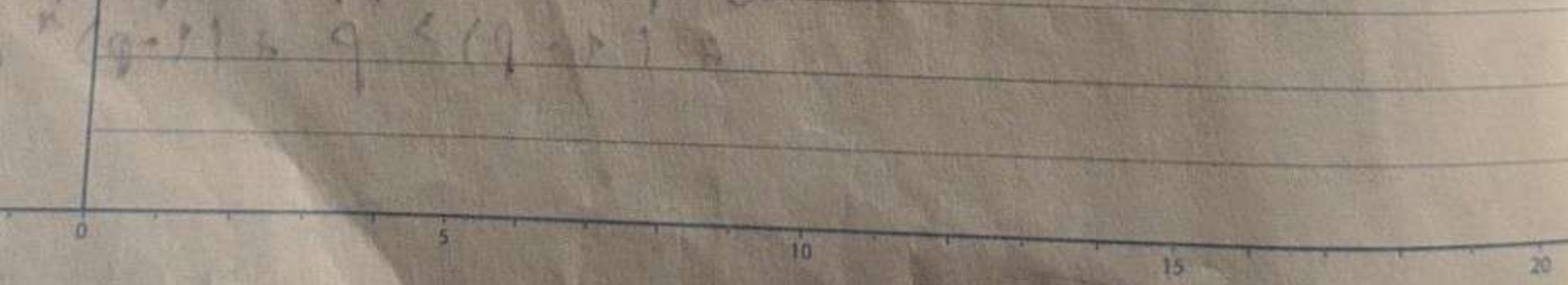
CHAPTER 4: Continuous random

!!) continuous ≈ measure (weight, height, temperature

(*) Definition: A continuous random variable is a random variable whose possible values includes in an interval of real number

(*) Probability density function
- $f(x) \geq 0$
- $\int_{-\infty}^{\infty} f(x)\,dx = 1$
- $P(a \leq X \leq b) = \int_a^b f(x)$
$= P(a < X < b)$





Ex 1: $f(x) = e^{-(x-7)}$ $x > 7$

a) $P(0 \leq x \leq 8)$

b) $P(x < 9)$

c) $P(x \geq 9)$

d) $P(x \geq 5)$

Ex 2: Find c such that $f(x)$ is a probability density function $f(x) = c(x^2 + 2x)$, $0 \leq x < 2$

Ex 1:

a) $P(0 \leq x \leq 8) = \int_0^8 f(x) = \int_7^8 e^{-(x-7)} = 0.6321$

b) $= \int_7^9 e^{-(x-7)}$

c) $P(x \geq 9) = 1 - P(x < 9)$

d) 1

Ex 2:

$$\int_0^2 f(x) = 1$$

$$c \cdot \left(\frac{1}{3} \cdot x^3 + x^2\right)\Big|_0^2 = 1$$

$$c\left(\frac{8}{3} + 4\right) = 1 \rightarrow c = 0.15$$

* $f(x) \geq 0 \ \forall x \Rightarrow c \geq 0$

* $\int_{-\infty}^{+\infty} f(x) = 1$

$$\rightarrow c \int_0^2 f(x) = 1$$

$$\rightarrow c \cdot \frac{20}{3} = 1$$

$$\rightarrow c = 0,15 \ (\geq 0)$$

Thứ
Ngày . . No.

(*) Probability function (cumulative)

$$F(a) = P(x \le a) = \int_{-\infty}^{a} f(x)\,d(x)$$

or $$F(X) = P(X \le x) = \int_{-\infty}^{x} f(t)\,dt$$

$$f(x) = F'(x)$$

Ex 1: $f(x) = e^{-(x-7)}$ $x > 7$

Find $F(x)$

$$F(x) = \begin{cases} \int e^{-(x-7)} = -e^{(x-7)} & x > 7 \\ 0 & x \le 7 \end{cases}$$

$$x > 7,\ F(x) = P(X < x) = \int_{-\infty}^{x} e^{-(t-7)}$$

$$\int_{7}^{x} e^{-(t-7)} = -e^{-(x-7)} + 1$$

Ex 2: Suppose the ~~cumulative~~ cumulative function of the random variable x is $F(x) = \begin{cases} 0, & x \le -1 \\ -x^2 + 1, & -1 \le x < 1 \\ 1 & x \ge 1 \end{cases}$

a) $P(x \le 0)$

b) $P(x \le -2)$

c) $P(x > 0,5)$

d) $P(-1 < x < 2)$

~~$f(x) = x$~~

0 5 10 15 20

a) $P(x \leq 0) = F(0) = 0^3 + 1 = 1$

b) $P(x < -2) = F(-2) = 0$

c) $P(x > 0,5) = 1 - P(x \leq 0,5) = 1 - F(0,5)$

$= 1 + 0,5^3 . 1 = 0,5^2$

d) $P(-1 < x < 2) = P(x<2) - P(x < -1)$

$= F(2) - F(-1)$

~~$= 1 - 0 = 1$~~

$= 1 + (1)^2 - 1 = 1$

⑧ Mean and Variance

[Mean

expected value

$$E(x) = \int_{-\infty}^{\infty} x f(x) dx = \mu$$

$$V(x) = \int_{-\infty}^{\infty} x^2 f(x) dx - \mu^2 = \sigma^2$$

$$\sigma = \sqrt{V(x)}$$

Ex 1: $f(x) = c(x^2 + 2x)$, $0 < x < 2$

Find $E(x)$, $V(x) = ?$

$c = 0.15$

$$E(x) = \int_0^2 x . 0.15 (x^2 + 2x) = 1.4 = \mu$$

$$V(x) = 0.15 \int_0^2 x^2 (x^2 + 2x) dx - 1.4^2 \mu^2$$

$$= 0.15 \int_0^2 x^4 + 2x^3 dx - 1.4^2$$

CHAPTER 4: CONTINUOUS DISTRIBUTIONS

① Continuous uniform distribution

$x \sim U[a, b]$

$f(x)$ = constant over $[a, b]$

$\Rightarrow \begin{cases} f(x) = 0, & \text{elsewhere} \\ f(x) = \frac{1}{b-a}, & a \le x \le b \end{cases}$

$\Rightarrow F(x) = \begin{cases} 0, & x < a \\ (x-a)/(b-a) & a \le x < b \\ 1, & x \ge b \end{cases}$

② Mean and variance

$E(x) = \frac{a+b}{2}$; $V(x) = \frac{(b-a)^2}{12}$

Ex 1: A plumber estimates that service call are uniformly distributed between 0.5 and 8 hour

a) What are mean and variance

b) Find prob call < 3 hours

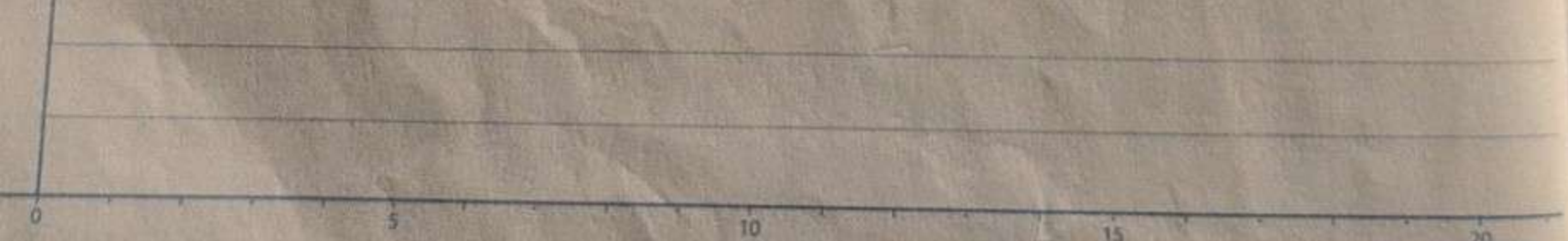
c) " " between 2-4 hours

d) " " more than 5 hours

e) " " more than 5 given that less than 7 hours

$E(x) = \frac{a+b}{2} = \frac{0.5+8}{2}$

$V(x) = \frac{(b-a)^2}{12} = \frac{7.5^2}{12}$



Thứ
Ngày No.

b) $P(x<3) = \int_{-0.5}^{3} \frac{1}{8-0.5}\,dx$

c) $P(2<x<9) = \int_{2}^{9} \frac{1}{8-0.5}\,dx$

d) $P(x>5) = \int_{5}^{\infty} \frac{1}{8-0.5}\,dx = \int_{5}^{8} \frac{1}{8-0.5}\,dx$

e) $P(5<x \mid x<7)$

$= \int_{5}^{7} \frac{1}{8-0.5}\,dx \;/\; \int_{0.5}^{7} \frac{1}{8-0.5}\,dx$

$E(ax+by) = aE(x) + bE(y)$

$E(ax+b) = aE(x) + b$

$V(ax+by) = a^2 V(x) + b^2 V(y)$

$V(ax+b) = a^2 V(x)$

$E(h(x)) = \sum_{i=1}^{n} h(x_i) P(x=x_i)$ (dis

$E(h(x)) = \int_{-\infty}^{\infty} h(x) f(x)\,dx$ (con

② Continuous normal distribution

- $f(x) \geq 0$
- bell-shaped
- symmetric

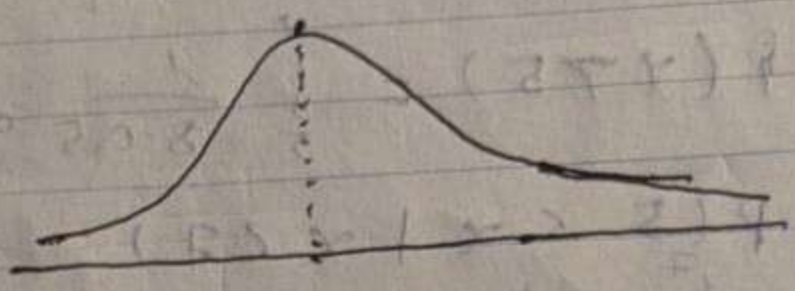
$f(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} \cdot e^{-1/2\left(\frac{x-\mu}{\sigma}\right)^2}$

KOKUYO

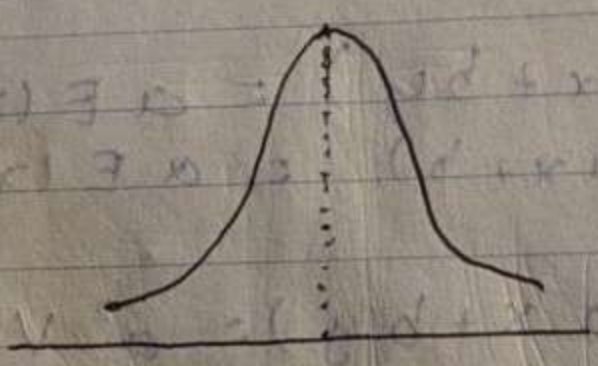
Thứ
Ngày
No.

* $X \sim N(\mu, \sigma^2)$

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{-\frac{1}{2}\left(\frac{x+\mu}{\sigma}\right)^2}$$



$$\downarrow Z = \frac{X-\mu}{\sigma}$$



$$P(X < a) = P\left(Z < \frac{a-\mu}{\sigma}\right) = \Phi\left(\frac{a-\mu}{\sigma}\right)$$

$\underbrace{z}$

$P(Z < a) = \Phi(a)$ (hàm tích lũy)

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

* $P(Z > a) = 1 - P(Z < a) = 1 - \Phi(a)$
* $P(a < Z < b) = P(Z < b) - P(Z < a)$
$= \Phi b - \Phi a$

Ex 1: $X \sim N(2100, 200^2)$ Find o

a) $P(x > 2000) = 1 - P(x < 2000)$

$= 1 - P\left(Z < \frac{2000 - 2100}{200}\right) =$

$= 1 - P(Z < -0.5) = 1 - \Phi(1)$

≈ 0.6189

b, $P(1700 < x < 2000)$

$= P\left(z < \frac{2000-2100}{200}\right) - P\left(z < \frac{1700-2100}{200}\right)$

$= P(z < -0.5) - P(z < -2)$

$= 0.285$

c) Find a such that $P(x > a) = 0.03$

$P(x > a) = 1 - P(x < a) = 0.97$

$\Rightarrow P\left(z < \frac{a-2100}{200}\right) = 0.97$

$\Rightarrow \Phi\left(\frac{a-2100}{200}\right) = \Phi\ 0.97$

4. $P(0.22 < z < a)$ $z \sim N(0, 1^2)$

$\Rightarrow P(z < a) - P(z < 0.22) = 0.24$

$\Rightarrow \Phi(a) = 0.24 + 0.59 = 0.83$

5. $X \sim N(16, 1.3^2)$

$P(x > 16.25) = 1 - P(x < 16.25)$

$= 1 - P\left(z < \frac{16.25-16}{1.3}\right)$

$= 1 - P(z < 0.192)$

$= 0.423$

3, Normal approximation to the Binomial and Poisson distribution

⊛ Normal approximation to Binomial

$n >>>>>$
$p <<<<$ ⇒ Normal $\xrightarrow{\text{approximation}}$ Binomial

$$Z = \frac{X - np}{\sqrt{npq}}$$

$\ominus P(X_{Binom} \leq a) = P(X_{NORM} \leq a + 0.5)$

$= P\left(Z \leq \frac{a + 0.5 - np}{\sqrt{npq}}\right)$

$\ominus P(X_{Binom} \geq a) = P(X_{NORM} \geq a - 0.5)$

$= 1 - P(X_{NORM} < a - 0.5)$

$= 1 - P\left(Z < \frac{a - 0.5 - np}{\sqrt{npq}}\right)$

⊛ ~~Normal~~ Normal approximation to Poisson

$X \sim$ Poisson (λ); Normal $\xrightarrow{\text{approximation}}$ Poisson

$$Z = \frac{X - \lambda}{\sqrt{\lambda}}$$

$P(X_{Poisson} \leq a) = P(X_{NORM} \leq a + 0.5)$

$\approx P\left(Z \leq \frac{a + 0.5 - \lambda}{\sqrt{\lambda}}\right)$

$P(X_{poisson} \geq a) = P(X_{NOM} \geq a + 0.5)$

$\approx P\left(Z \geq \frac{a - 0.5 - \lambda}{\sqrt{\lambda}}\right)$

Thứ
Ngày No.

Ex 11 :

$p = 0.7$, $n = 800$

$P(X_{BINORM} > 579) = P(X_{BINORM} \geq 580)$

$= P(X_{NORM} \geq 579.5) = 1 - P(X_{NORM} \leq 579.5)$

$= 1 - P(Z < \frac{579.5 - 800 \cdot 0.7}{\sqrt{800 \cdot 0.7 \cdot 0.3}})$

$= 1 - P(Z < 1.509)$

Ex 12 :

$p = 0.03$, $n = 800$

a) $P(X_{BINORM} > 30) = P(X_{BINORM} \geq 31)$

$= P(X_{NORM} \geq 30.5)$

$= 1 - P(X_{NORM} \leq 30.5)$

$= 1 - P(Z < \frac{30.5 - 800 \cdot 0.03}{\sqrt{800 \cdot 0.03 \cdot 0.97}})$

b) $P(20 < X_{BINORM} \leq 30) = P(21 \leq X_{BIN} \leq 29)$

$= P(20.5 \leq X_{NORM} \leq 29.5)$

$= P(X_{NORM} \leq 29.5) - P(X_{NORM} \leq 20.5)$

Ex 13 :

$E(X) = \lambda = 1000$

$P(X_{POISSON} \leq 950) = P(950 \leq P_{Poi} \leq 950)$

$= P(949.5 \leq X_{NORM} \leq 950.5)$

$= P(X_{NORM} \leq 950.5) - P(X_{NORM} < 949.5)$

5 10 15 20

KOKUYO

Ex 14:

$E(x) = \lambda = 9.6 > 5 \Rightarrow$ using Normal approximation

$P(X_{POISSON} > 10) = P(X_{POISSON} \geq 11)$

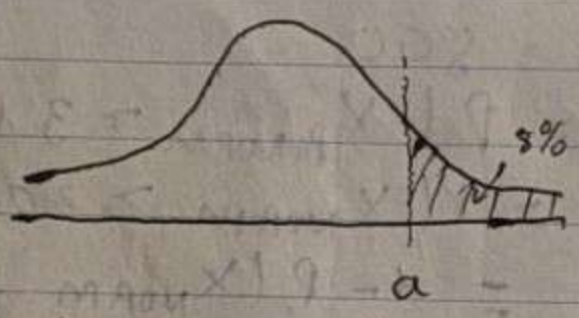
$= 1 - P\left(Z < \frac{10.5 - 9.6}{\sqrt{9.6}}\right)$

Ex 7:

$E(x) = \mu = 8.21, \quad \sigma = 2.19$

Find a, $P(X > a) = 8\%$

$P(X < a) = 0.92$

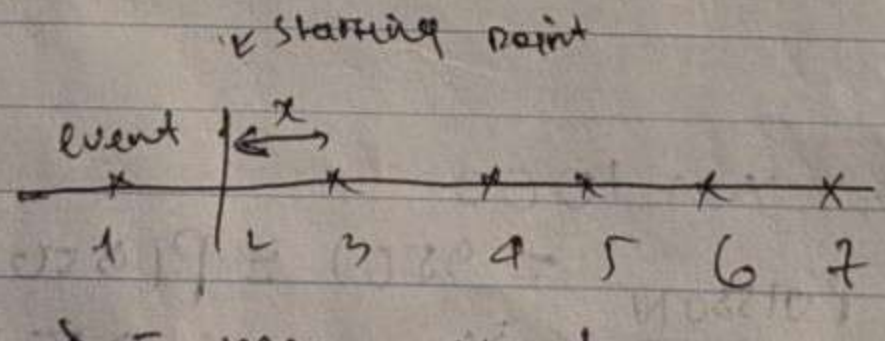


$P\left(Z < \frac{a - 8.21}{2.19}\right) = 0.92$

$\Phi\left(\frac{a - 8.21}{2.19}\right) = \Phi(NORM.INV(0.92))$

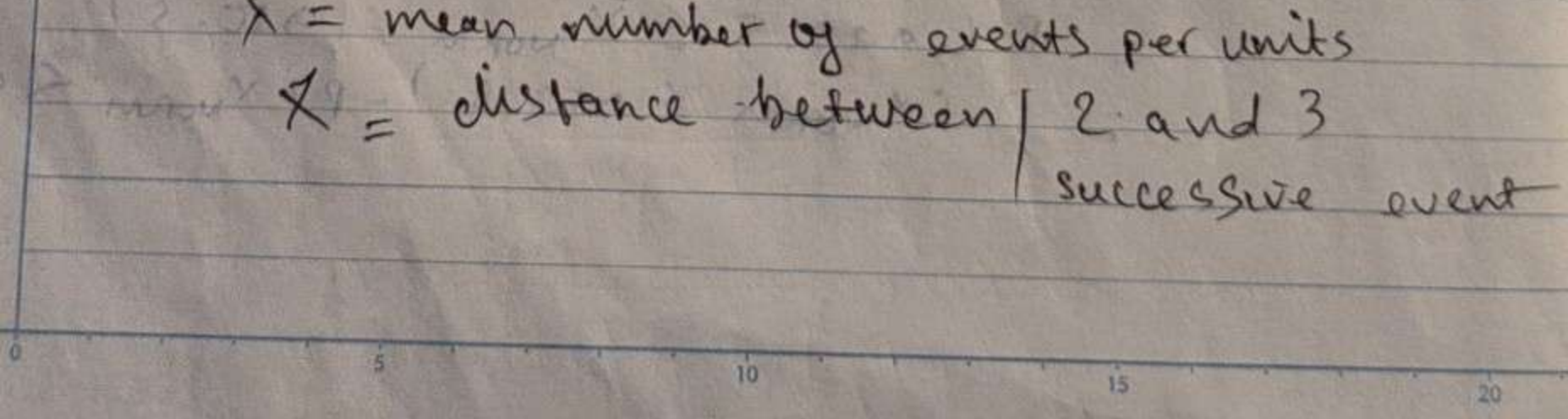
$\rightarrow \frac{a - 8.21}{2.19} = NORM.S.INV(0.92)$

4) Exponential distribution



λ = mean number of events per units

X = distance between | 2 and 3 successive event



$P(X > x) = e^{-\lambda x} \quad (x \geq 0)$

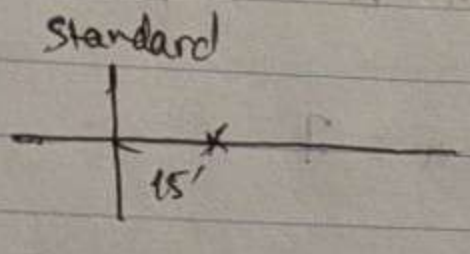
$F(x) = P(X \leq x) = 1 - e^{-\lambda x}$

* Mean and variance

$E(X) = \frac{1}{\lambda}$

$V(X) = \frac{1}{\lambda^2}$

Ex 15, $E(X) = 9 = \frac{1}{\lambda} \rightarrow \lambda = \frac{1}{9}$



$P(X > 15) = e^{-1/9 \cdot 15}$

$=$

Ex 16, $E(X) = 3.5 \rightarrow \lambda = 1/3.5$

$P(X \geq 6) = e^{-\lambda 6} = e^{-1/3.5 \times 6}$

Ex 17, $\lambda = 15/h$

$P(X \leq 6') = P(X \leq 1/10\,(h)) = 1 - e^{-15 \cdot \frac{1}{10}}$

CHAPTER 6: Descriptive Statistics

Take a sample of Population (data) $x_1, x_2, \ldots, x_n$

From a population N

* Definition

① sample mean $\bar{x} = \frac{\Sigma x_i}{n}$

② Sample median (sorted data)

$L = \frac{n+1}{2} \rightarrow \text{median} = \frac{x_{\lceil \frac{n+1}{2} \rceil} + x_{\lfloor \frac{n+1}{2} \rfloor}}{2}$

Ex 1: 8 7 8 8 7 7

→ 7 7 7 8 8 8 (sorted), $n = 6$

$L = \frac{6+1}{2} = 3.5$

→ median $= \frac{x_3 + x_4}{2} = \frac{7+8}{2}$

Ex 2: 1 2 3 7 8 9 10

$L = \frac{7+1}{2} = 4$

→ median $= x_4 = 7$

③ Sample mode

Ex 1:

1, 8 7 8 8 7 7

mode: 7, 8

2, 5 6 8 8 10 10

mode: 8, 10

3, 1 2 3 7 8 9 10

mode: #

④ Variance

1, Sample range

$\Rightarrow r = \max(x_i) - \min(x_i)$

2, Variance

$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$ ②

$s^2 = \frac{\sum x_i^2 - n\bar{x}^2}{n-1}$ ③

Standard deviation : $s = \sqrt{s^2}$

Ex: 32 41 47 50 56 58 61
62 68 75 82 85 90 92 98

Find sample mean and standard deviation

$\bar{x} = 66.96$ ~~65.94~~

~~$s \approx 18.7832$~~

$s = 19.4$

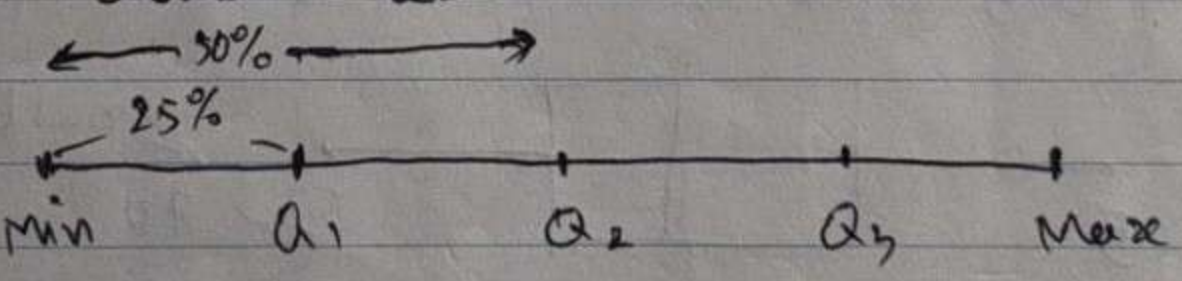
⑤ Stem and leaf diagram

Ex1: 55 115 225 290 330
335 385 900 905 905

stem	leaf
5	5
11	5
22	5
29	0
33	0 5
38	5
40	0 5 5

(!) X.Y : stem X, leaf Y

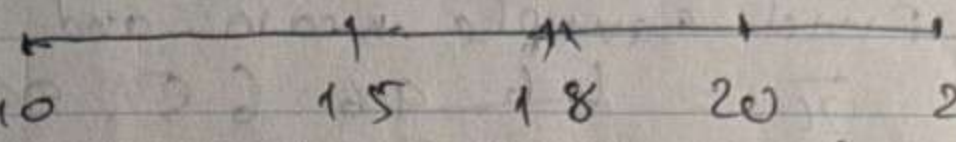
⑥ Quartiles



⑦ Interquartile range IQR $= Q_3 - Q_1$

Ex: 55, 52, 52, 52, 49, 71, 67, 55
Find the sample quartiles and interquatile range

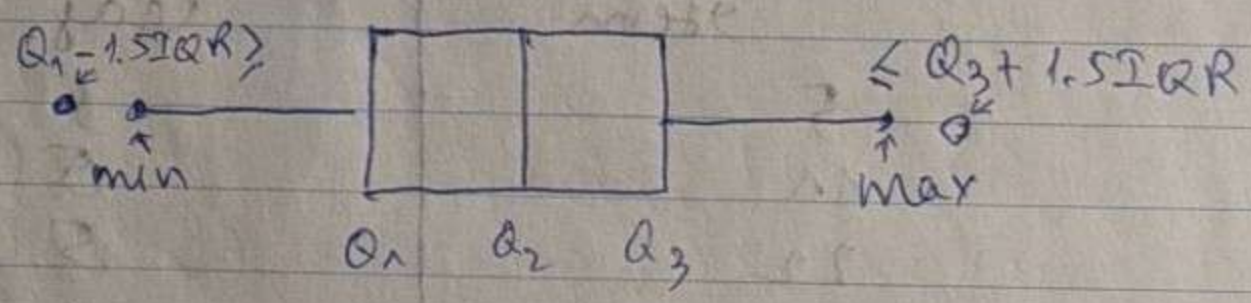
Ex 3:



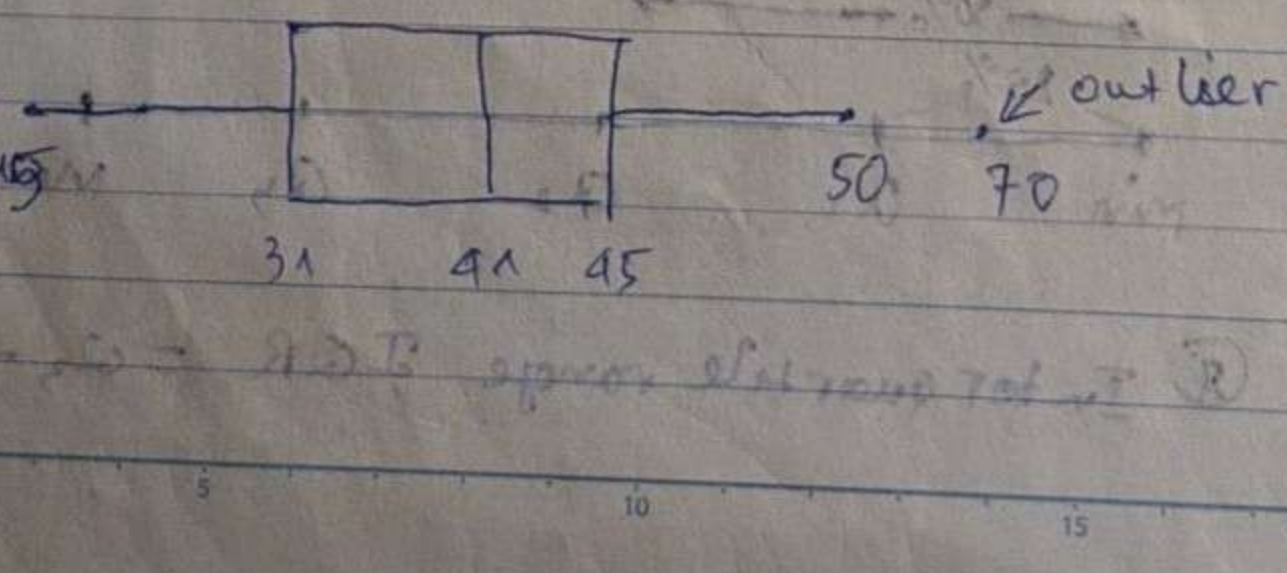
How many of the following statements are correct

i) About 50% of the data lies above 15
ii) IQR = 15
iii) The number 10 must be in the data set

8. Box plot



Ex: 15, 20, 31, 31, 32, 40, 41, 41, 42, 43, 45, 45, 50, 70
Draw a box plot for this data
min = 15, $Q_1 = 31$, $Q_2 = 41$, $Q_3 = 45$, Max = 70
$IQR = Q_3 - Q_1 = 14$
$Q_1 - 1.5\ IQR = 10$
$Q_3 + 1.5\ IQR = 66$



Thứ
Ngày . . No.

box plot

Ex2 : Construct the ~~data set~~ for the data set
100, 113, 115, 116, 118, 123, 125, 136, 136,
138, 140, 142, 145, 146, 149, 150, 156, 162, 163

Ex3:

Stem	Leaf
0	5 5 6
1	1 4 6 2 3
2	0 2 1

Ex2:

$Q_1 = 117$, $Q_2 = 137$, $Q_3 = 197.5$

$IQR = 80.5$

Min = $117 - 1.5 \times 80.5 = 71.25$

Max = $117 + 1.5 \times 80.5 = 193.25$

No outlier

Ex3:

$Q_3 = 20$

(9) Frequency distribution (both ~~circle~~ disc and cont)

Ex: 1 1 2 2 2 3 3 4 5 6 7 7 8

x	Freq
1	2
2	3
3	2
4	1
5	1
6	1

0 5 10 15 20

KOKUYO

Thứ
Ngày . . No.

=7 data : chvide : 5-20 [classes, interval
bins

X	Freq
[1-2]	5
[3-4]	3
[5-6]	2
[7-8]	3

Ex 2:

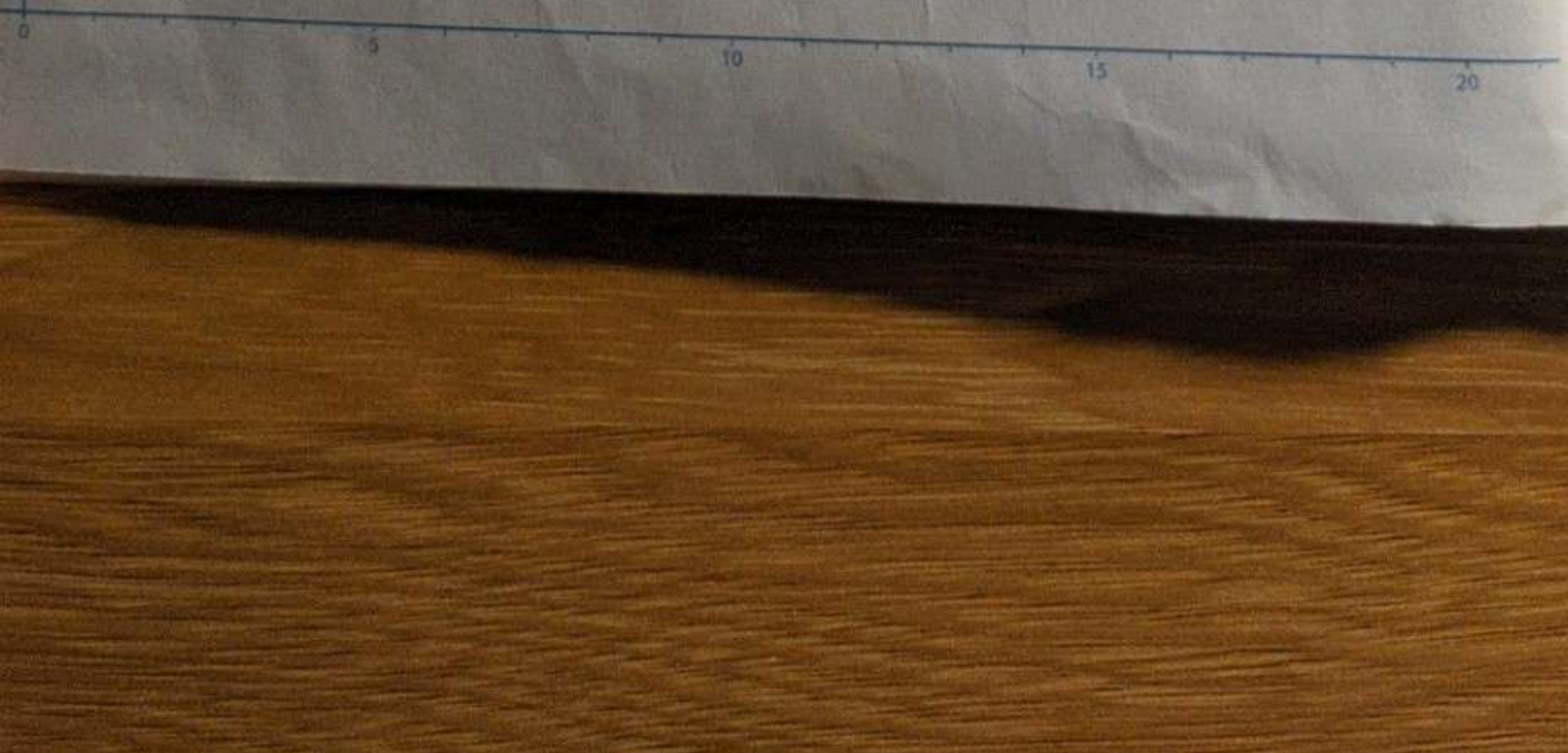
X	Freq	Cumulative Freq	Relative Freq	Cum Rel Freq
[0,2)	2	2	2/30	2/30
[2,4)	5	7	5/30	7/30
[4,6)	8	15	8/30	15/30
[6,8)	11	26	11/30	26/30
[8,10)	4	30	4/30	1

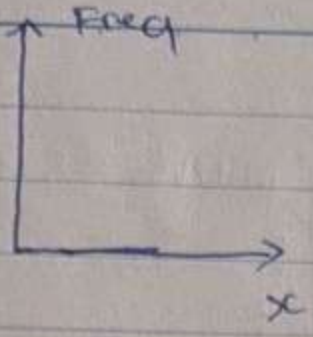
Ex 3: The age of 25 member : take part in club

16 - 20	20%
21 - 30	28%
31 - 40	36%
41 - 50	16%

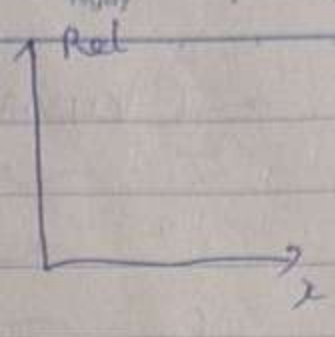
How many members in club that their age are not exceed 40?

⑩ Histogram





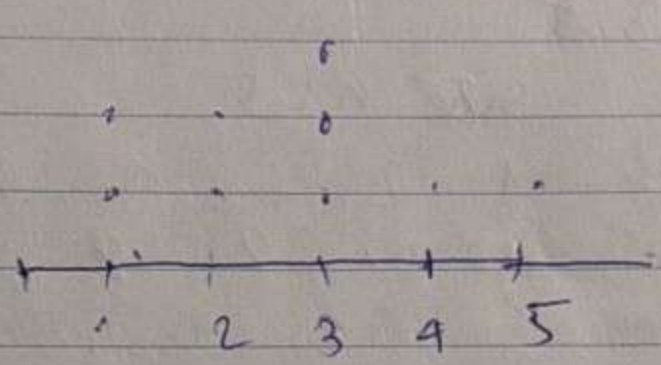
Histogram



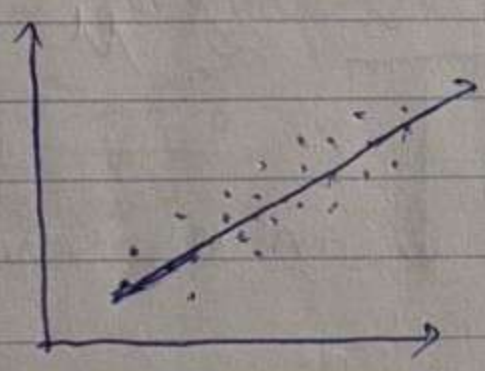
Relative Histogram

(11) Dot plot

Ex: 1 1 2 2 3 3 3 4 5



(12) Scatter plot



CHAPTER 7: Sampling distribution and part estimator

Big picture :

KOKUYO

Thứ
Ngày . . No.

* Claim

$\bar{x} \rightsquigarrow \mu$ $\quad$ $\bar{x}$: point estimator for μ

$s^2 \rightsquigarrow \sigma^2$ $\quad$ s^2 : _____ σ^2

$\hat{p} \rightsquigarrow p$ $\quad$ $\hat{p}$: _____ p

① Point estimator for μ

X	1	2	3	4	5	6
P(X)	1/6	1/6	1/6	1/6	1/6	1/6

$\mu = 3.5$

sample : size $n = 5$

$c = 1$: 1 2 3 4 5 $\rightarrow \bar{x}_1 = 3$

$c = 2$: 1 1 2 2 6 $\rightarrow \bar{x}_2 = 2.4$

....

$c = n$: $\rightarrow \bar{x}_n = ...$

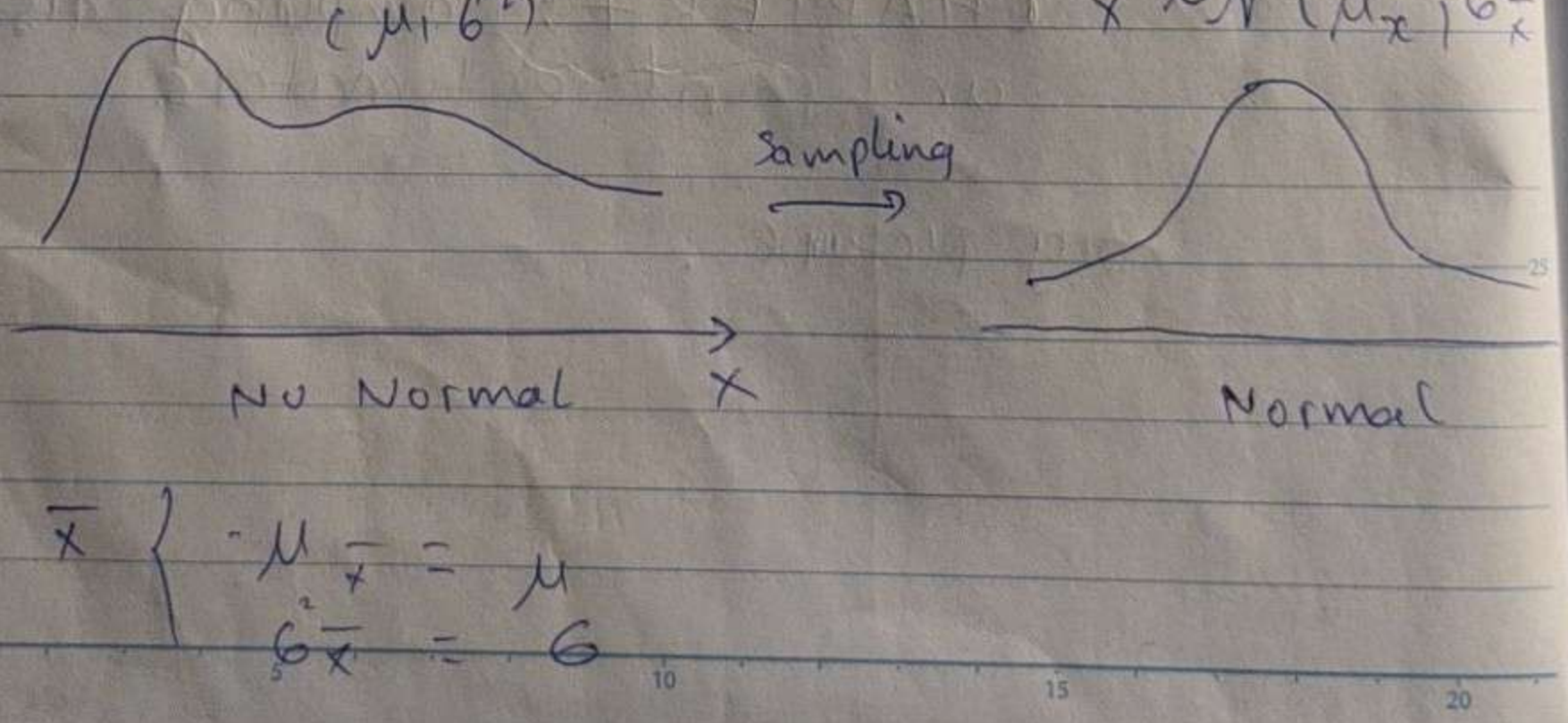
$$\frac{\bar{x}_1 + \bar{x}_2 + .. + \bar{x}_n}{n} \approx \mu$$

Remark : The mean of sample : $\mu_{\bar{x}} = \mu$

point estimator for μ $\bar{X} \sim \mu$

↓ unbiased estimator

(μ, σ^2) $\quad$ $\bar{x} \sim N(\mu_{\bar{x}}, \sigma^2_{\bar{x}})$



$\bar{x}$ { $\mu_{\bar{x}} = \mu$; $\sigma^2_{\bar{x}} = \sigma$

5 10 15 20

Standard error : $6_{\bar{x}} = \frac{6}{\sqrt{n}}$

* Central limit theorem

[x : normal
$n \geq 30$ $\rightarrow \bar{x} \sim N(\mu, \frac{6^2}{n})$ $\xrightarrow{z = \frac{\bar{x}-\mu}{6/\sqrt{n}}}$

sample — normal

$Z \sim N(0, 1)$

standar normal

- Theorem [CLT for two population]

If we have 2 independent population with parameters $(\mu_1, 6_1^2)$ and $(\mu_2, 6_2^2)$, and if $\bar{X}_1$ and $\bar{X}_2$ are the same means of 2 independent random samples of size n_1 and n_2 from these populations, then the sampling distributions of

$$\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{6_1^2}{n_1} + \frac{6_2^2}{n_2}}}$$

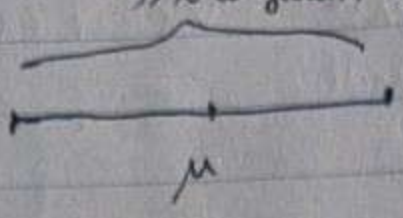
is approximately standard normal for large n_1, n_2. It is exactly standard normal if the 2 populations are normal

$$\bar{X}_1 - \bar{X}_2 \sim N(\mu_1 - \mu_2 ; \frac{6_1^2}{n_1} + \frac{6_2^2}{n_2})$$

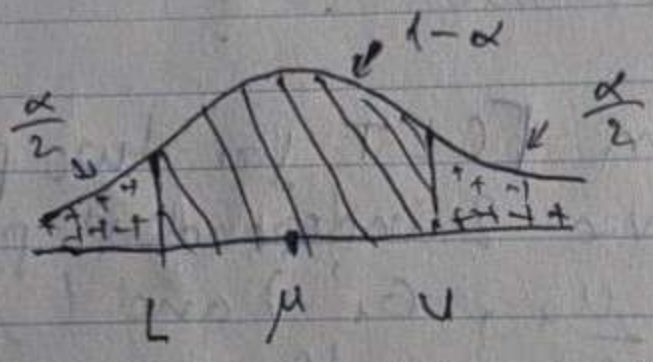
CHAPTER 8: CONFIDENCE INTERVAL

95% confident level

Ex: $\bar{X}$,



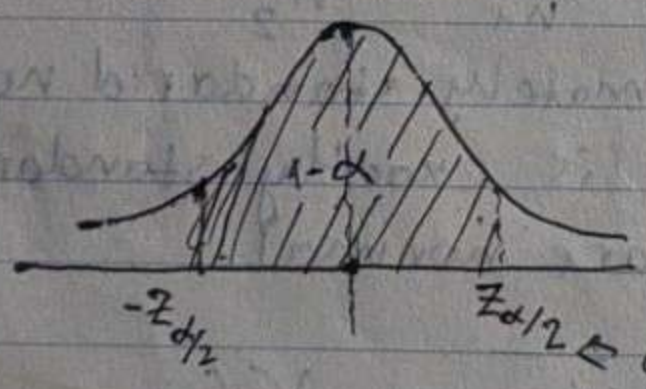
① Confidence level for mean (σ : known)



$$P(L \leq \mu \leq U) = 1 - \alpha$$

$1-\alpha$: confidence / coefficient level

(Sample n) $\Rightarrow Z = \dfrac{\bar{X} - \mu}{\sigma / \sqrt{n}}$



$$P(-Z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < Z_{\alpha/2}) = 1 - \alpha$$

$$= P\left(\bar{X} - Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}\right) = 1 -$$

Thứ
Ngày . . No

* Margin of error

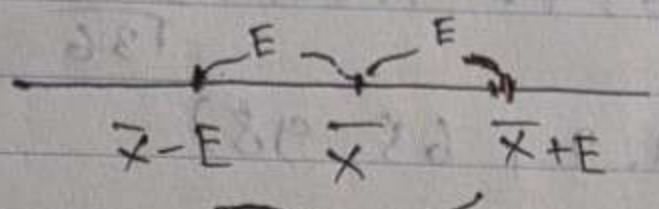
$$E = z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

$\Rightarrow \bar{x} - E < \mu < \bar{x} + E$

$\Rightarrow (1-\alpha) \cdot 100\%$ confidence interval for mean

$(\bar{x} - E, \bar{x} + E)$

(2)



(3) $E \leq a, 6, 1-\alpha$

$\Rightarrow$ How large of sample mean?

↳ find $n = ?$

$$E = z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \Rightarrow n = \left\lceil \left(\frac{\sigma \cdot z_{\alpha/2}}{E} \right)^2 \right\rceil$$

= critical value / standard error

$$z_{\alpha/2} = -NORM.S.INV\left(\frac{\alpha}{2}\right) = NORM.S.INV\left(1-\frac{\alpha}{2}\right)$$

Confidence level	$z_{\alpha/2}$
90%	1.64
95%	1.96
98%	2.33
99%	2.58

0 5 10 15 20

KOKUYO

Ex 1:

$n = 36$

$\bar{X} = 65.3, \quad 6 = 2.5$

$1 - \alpha = 0.9 \Rightarrow \frac{\alpha}{2} = 0.05 \Rightarrow z_{0.05} = 1.64$

$\Rightarrow$ 90% CI for mean is $(\bar{X} \pm E)$

$$\left(65.3 - 1.64 \cdot \frac{2.5}{\sqrt{36}}, \; 65.3 + 1.64 \cdot \frac{2.5}{\sqrt{36}}\right)$$

$$= (64.2; \; 65.98)$$

90% confidence that the true mean is between 64.2 and 65.8

Ex 2:

$1 - \alpha = 0.98 \Rightarrow \frac{\alpha}{2} = 0.01 \Rightarrow z_{0.01} = 2.33$

$6 = 6 \qquad \mu = \frac{2}{4}$

$$n = \left\lceil \left(\frac{2.33 \cdot 6}{4}\right)^2 \right\rceil = 13$$

Ex 3:

$\mu_{\bar{x}} = 21000, \; 6 = 1000, \; z_{0.05} = 1.96, \; n = 64$

$$2E = 2 \cdot z_{\alpha/2} \frac{6}{\sqrt{n}} = 2 \cdot 1.96 \cdot \frac{1000}{\sqrt{64}} = 490$$

Ex 4:

$z_{0.05} = 1.96$

$E \leq 500$

$6 = 4000$

$$n = \left\lceil \left(\frac{4000 \cdot 1.645}{500}\right)^2 \right\rceil = 17$$

Ex5:

$z_{0.005} = 2.58$

$E = 8$

$\sigma = 21$

$n = \lceil \left(\frac{21 \cdot 2.58}{8}\right)^2 \rceil = 4$

* $E \downarrow (=) [\ n \uparrow$
  $z_{\alpha/2} \downarrow (=) \alpha/2 \uparrow (=) 1-\alpha \downarrow$
* One-sided confidence bounds
  $(1-\alpha) \cdot 100\%$ lower-confidence bound is $\bar{x} - E$
  $(1-\alpha) \cdot 100\%$ upper-confidence bound is $\bar{x} + E$
  E = critical value . standard error
  $E = z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$

Ex9: $\sigma = 0.01$ $1-\alpha = 0.98 \rightarrow \alpha = 0.02 \downarrow$ $z_{0.02} = 2.05$

$n = 15$

$\bar{x} = 1.5$

→ 98% lower confidence bound: $\bar{x} - E = 1.5 - \frac{2.05 \cdot 0.01}{\sqrt{15}} \approx 1.995$

Ex11:

$\bar{x} = 82$ $n = 30$

$\sigma = 12.2$

a) $1 - \alpha = 0.98 + \frac{0.1}{10} \rightarrow \alpha = 0.01 \rightarrow z_{0.01} = 2.58$

90% CI → $\left(\bar{x} - \frac{z \cdot \sigma}{\sqrt{n}} ; \frac{\bar{x} + z \cdot \sigma}{\sqrt{n}}\right)$

$= (76.25 ; 87.7467)$

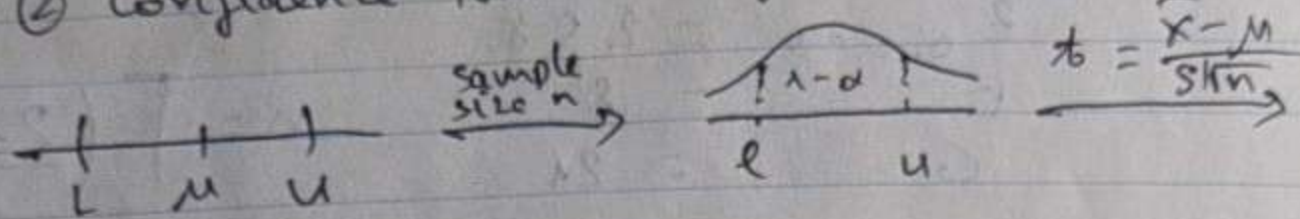
b) $z_{0.02} = 2.05$

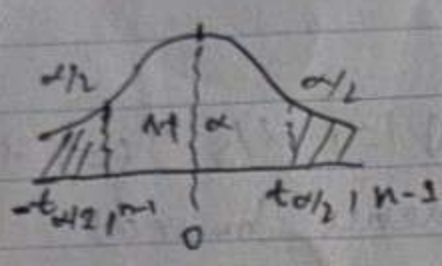
→ 98% upper-confidence bound

$= \bar{x} + \frac{z_{0.02} \cdot \sigma}{\sqrt{n}} \approx 86.56$

Thứ
Ngày . . No.

② Confidence interval for mean (σ : unknown)





S = (sample) standard deviation $= \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}}$

un biased estimator

degree of freedom
$df = n-1$

$$E = t_{\alpha/2,\,n-1} \cdot \frac{S}{\sqrt{n}}$$

Ex5:

$n = 12$

$\bar{x} = 916.8$

$s = 15.2$

$1-\alpha = 0.99 \Rightarrow \alpha = 0.01 \Rightarrow t_{0.005,11} = 3.106$

$(\bar{x} - E,\ \bar{x} + E) = (303,171,\ 330,4287)$

Ex6:

$\bar{x} = 2259,917$

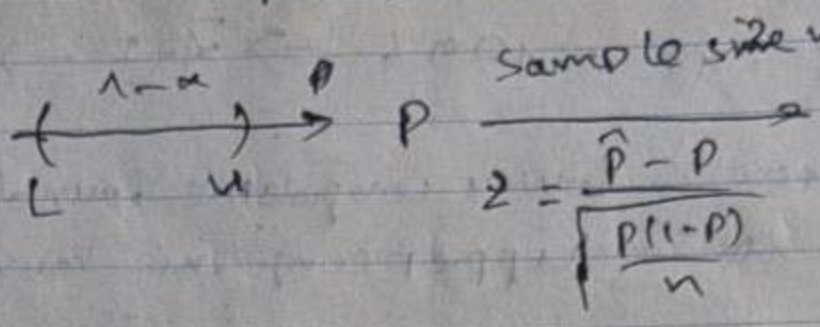
$s = 35,57$

$n = 12$

as $1-\alpha = 0.95 \Rightarrow t_{0.025,11} = 2,201$

$\Rightarrow (\bar{x} - E,\ \bar{x} + E) = (2239,316,\ 2282,517)$

when $\begin{cases} np \geq 5 \\ np(1-p) \geq 5 \end{cases} \Rightarrow \hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$



$$\rightarrow P\left(-z_{\alpha/2} < \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} < z_{\alpha/2}\right)$$

$$\rightarrow P\left(\hat{p} - z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} < p < \hat{p} + z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}}\right)$$

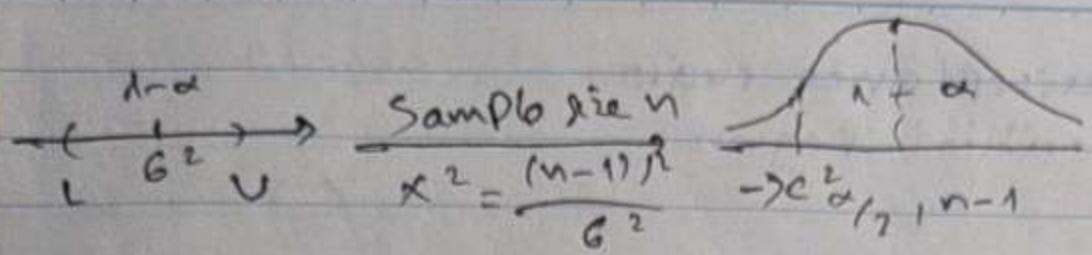
$$E = z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

* If p is unknown, $\hat{p} = \hat{p}_{max} = 0.5$

* One sided confidence bound for proportion
$(1-\alpha)$ 100% lower-confidence bound for proportion is $\hat{p} - E$
$(1-\alpha)$ 100% upper-confidence bound for proportion is $\hat{p} + E$
E = critical value standar error

$$E = z_{\alpha} \cdot \sqrt{\frac{\hat{p}(\hat{p}-1)}{n}}$$

mterval for
④ Confidence Vvarance.
$S^2 \rightsquigarrow \sigma^2$



$$\rightarrow 1-\alpha = P\left(\frac{(n-1)S^2}{x^2_{\alpha/2,n-1}} < \sigma^2 < \frac{(n-1)S^2}{x^2_{1-\alpha/2,n-1}}\right)$$

$\Rightarrow (1-\alpha).100\%$ CI for variance is

$$\left(\frac{(n-1)S^2}{x^2_{\alpha/2,n-1}} , \frac{(n-1)S^2}{x^2_{1-\alpha/2,n-1}}\right) = N(A,B)$$

$(1-\alpha).100\%$ CI for standard deviation is $N(\sqrt{A}, \sqrt{B})$

CHAPTER 9: Test of hypothesis for a single sample

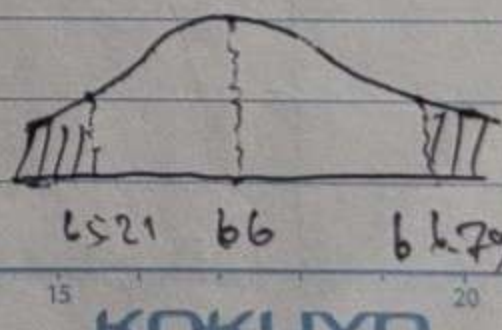
Ex: 95% CI for mean is (65.21 , 66.79)

95% confident that the true mean is between 65.21 and 66.79

① claim : $\mu = 60$ → reject claim

② claim : $\mu = 66$ → maybe → fail to reject claim

③ claim : $\mu \neq 66$ → reject claim



∈ ▨: reject

∈ □: fail to reject

Thứ
Ngày No.

□ : acceptance region
▨ : critical region

Def :
1, Statistical hypothesis is a statement about the parameter (μ, σ^2, P) of one or more population

2, claim => { H_0 : null hypothesis (contain "=")
H_1 : alternative hypothesis (not contain "=")

Ex 1: 1, claim : $\mu = 60$ → { $H_0 = \mu = 60$
$H_1 = \mu \neq 60$

2, claim $\mu \neq 66$ → { $H_0 = \mu = 66$
$H_1 = \mu \neq 66$

3, claim $\mu > 66$ → { $H_0 = \mu \leq 66$
$H_2 = \mu > 66$

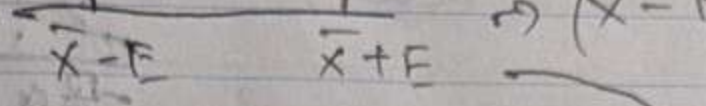
4,

Ex: claim $\mu = 66$ $\Rightarrow$ $\begin{cases} H_0: \mu = 66 \\ H_1: \mu \neq 66 \end{cases}$

(65.21 ; 66.79)

Sample $\Rightarrow \bar{x}$

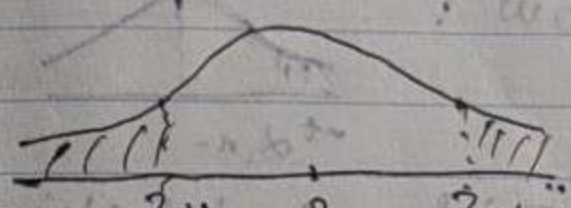
Chap 8



$\Rightarrow (\bar{x}-E, \bar{x}+E)$

$\rightarrow \begin{cases} \mu \in \ldots \\ \mu \notin \ldots \end{cases} \rightarrow$ conclusion

$$z_0 = \frac{\bar{x}-\mu}{\sigma/\sqrt{n}}$$

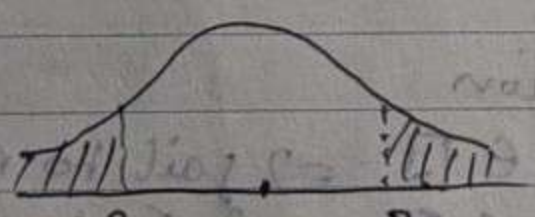


$\rightarrow \begin{cases} z_0 \in \square : \text{fail to reject } H_0 \\ z_0 \in \blacksquare : \text{reject } H_0 \Rightarrow \text{reject claim} \end{cases}$

Ex: claim: $p > 0.4$

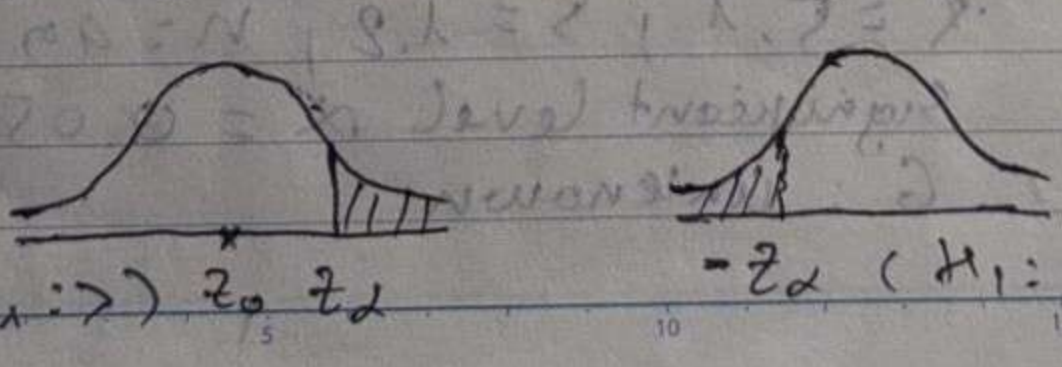
$\Rightarrow \begin{cases} H_0: p \leq 0.4 \leftrightarrow p = 0.4 \\ H_1: p > 0.4 \end{cases}$

$$z_0 = \frac{\hat{p}-p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$



$z_0 \in \square$: $\rightarrow$ fail to reject $H_0 \rightarrow$ reject claim

$z_0 \in \blacksquare$: $\rightarrow$ reject $H_0 \rightarrow$ fail to reject claim



KOKUYO

Thứ
Ngày . . No

* Test of a hypothesis (right tail test) (two tail test)

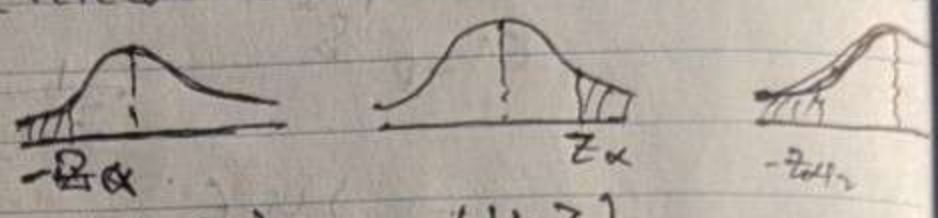
Step 1: claim $\mu < \mu_0$ $\mu \leq \mu_0$ $\mu = \mu_0$

Step 2: $H_0: \mu \geq \mu_0$ $H_0: \mu \leq \mu_0$ $H_0: \mu = \mu_0$

$H_1: \mu < \mu_0$ $H_1: \mu > \mu_0$ $H_1: \mu \neq \mu_0$

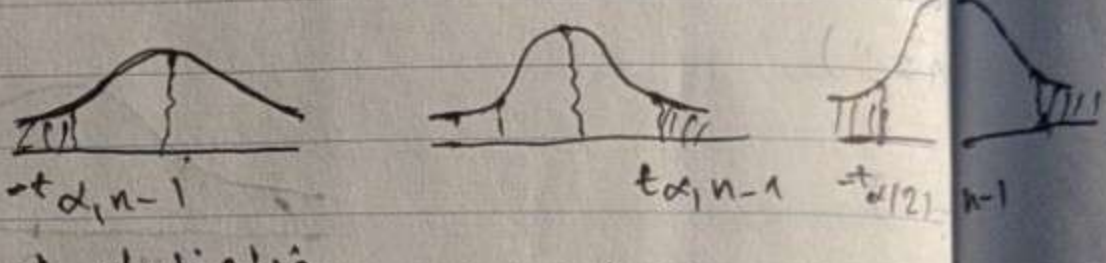
Step 3: Find critical value

σ: Know:



$(H_1 <)$ $(H_1 >)$

σ: unknow:



Step 4: Find test statistic

test statistic

- For mean: σ: known: $z_0 = \dfrac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$

[normal
$n \geq 30$

σ: unknown: $t_0 = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$

- For proportion $z_0 = \dfrac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$

Step 5: Conclusion

test statistic $\in$ [ ] => fail to reject H_0

$\in$ [ ] => reject H_0

(!) ~~Note~~

Ex3:

$\bar{x} = 5.1$, $s = 1.2$, $n = 90$

Significant level $\alpha = 0.05$

σ: unknown

claim $\mu \geq 9.5 \rightarrow \begin{cases} H_0 = \mu \leq 9.5 \\ H_1 = \mu > 9.5 \end{cases}$

(right-tailed test)

critical value: $t_{\alpha/n-1} = t_{0.05,48} = 1.677$

test statistic: $t_0 = \dfrac{\bar{x} - \mu_0}{\frac{s}{\sqrt{n}}} = \dfrac{5.1 - 9.5}{\frac{4.2}{7}} = -3$

$t_0 \in \square \Rightarrow$ reject H_0

→ fail to reject H_0

④ claim $p = 0.6$ $\begin{cases} H_0 = p = 0.6 \\ H_1 = p \neq 0.6 \end{cases}$

$n = 140$, $\hat{p} = \frac{x}{n} = \frac{75}{140} = 0.53$

$\alpha = 0.1$

$\rightarrow z_{\alpha/2} = z_{0.05} = 1.645$

and $z_0 = \dfrac{\hat{p} - p_0}{\sqrt{\frac{p(1-p_0)}{n}}} = -1.55 \in \square$

→ fail to reject H_0

CI (65.21, 66.79)

↓

claim

↓

reject claim → error

	H_0 true	H_0 false
Fail to reject H_0	✓ $(1-\alpha)$	Error type II (β) $(1-\beta)$
Reject H_0	Error type I (α)	✓ $(1-\beta)$

$1-\alpha$: confidence level

$P(\text{Reject } H_0 \mid H_0 \text{ true}) = \alpha$: significant level

$P(\text{fail to reject } H_0 \mid H_0 \text{ false}) = 1-\beta$: the power of statistical test

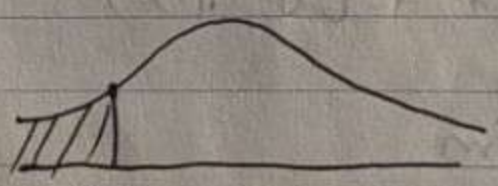
Fix $n \rightarrow \begin{cases} \alpha : \text{fix} \\ \beta : \downarrow \end{cases}$

Ex5: Chon $\mu < 150$

$\Rightarrow \begin{cases} H_0 : \mu \geq 150 \ (\mu = 150) \\ H_1 : \mu < 150 \end{cases}$

$\sigma = 16.7, \ \bar{x} = 153.36, \ \alpha = 0.01, \ n = 11$

$$z_0 = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{153.36 - 150}{16.7/\sqrt{11}} = 0.67$$



$-z_{0.01} = -2.33$

$(H_1 <)$

$\rightarrow z_0 \in \square \rightarrow$ fail to reject H_0

$\rightarrow$ reject claim

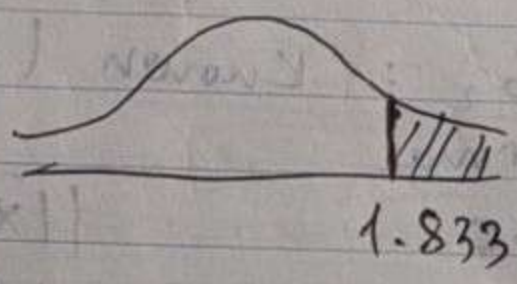
Thứ
Ngày . . No.

Ex6: $\bar{X} = 537.1$, $\alpha = 0.05$, $t_{0.05;9} = 1.833$

$S = 20.7$,

$\mu > 520 \rightarrow \begin{cases} H_0: \mu \leq 520 \\ H_1: \mu > 520 \end{cases}$

$$t_0 = \frac{537.1 - 520}{20.7/\sqrt{9}} = 2.481$$



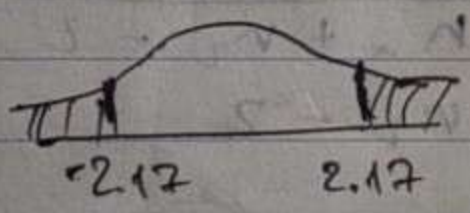
$\rightarrow \in$ ▨ $\rightarrow$ reject H_0

$\rightarrow$ fail to reject claim

Ex7: $\hat{p} = 0.06$, $n = 200$

$\begin{cases} H_0 \div \mu = 3\% = \alpha \rightarrow z_{\alpha/2} = 2.17 \\ H_1: \mu \neq 3\% \end{cases}$

$$z_0 = \frac{0.06 - 0.03}{\sqrt{\frac{0.03(1-0.03)}{200}}} = 2.487$$



$\rightarrow \in$ ▨ $\rightarrow$ reject H_0

$\rightarrow$ reject claim

KOKUYO

Thứ
Ngày
No.

CHAPTER 10: Statistical Inference for two sample

For mean $(\mu_1 - \mu_2)$

[X, Y: normal, $n_1, n_2 \geq 30$] $\Rightarrow \bar{X} \sim N(\mu_1, \frac{\sigma_1^2}{n_1})$, $\bar{Y} \sim N(\mu_2, \frac{\sigma_2^2}{n_2})$

$\Rightarrow \bar{X} - \bar{Y} \sim (\mu_1 - \mu_2, \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})$

Table for $\mu_1 - \mu_2$

① σ_1, σ_2: Known $((\bar{X} + \bar{Y}) \pm E)$ CI $((\bar{X} - \bar{Y}) \pm E)$

[X, Y normal, $n_1, n_2 \geq 30$] $\Rightarrow$ use z

E = critical value · Standard error

$$E = z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} \qquad z_0 = \frac{(\bar{X} - \bar{Y}) - (\mu_0 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

② σ_1, σ_2: unknow

a) assume: $\sigma_1^2 = \sigma_2^2$

$\Rightarrow$ pooled variance

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

$$df = n_1 + n_2 - 2$$

$((\bar{x} - \bar{y}) \pm E)$

$$E = t_{\alpha, df} \sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

$$t_0 = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}}$$

$$1 < \frac{s_{large}^2}{s_{small}^2} < 2 \rightarrow \sigma_1^2 = \sigma_2^2$$

b) not assume : $\sigma_1^2 = \sigma_2^2$

~~df =~~

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{(s_1/n_1)^2}{n_1 - 1} + \frac{(s_2/n_2)^2}{n_2 - 1}}$$

$$E = \alpha/2, df \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$t_0 = \frac{(\bar{x} - \bar{y}) - \Delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Ex:

$\cdot \mu_1 - \mu_2 > 0$

$H_0 : \mu_1 - \mu_2 \leq 0$

$H_1 : \mu_1 - \mu_2 > 0$

test statistic $t_0 = \frac{\bar{x}_1 - \bar{x}_2 - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$

$= -0.$

$$S_p^2 = \frac{(n_2 - 1)s_1^2 + (n_1 - 1)s_2^2}{n_1 + n_2 - 2} = 6.97875$$

$df = n_1 + n_2 - 2 = 29$

$$t_0 = \frac{(\bar{x} - \bar{y}) - (\mu_1 - \mu_2)}{\sqrt{\frac{S_p^2}{n_1} + \frac{\sum p^2}{n_2}}} = -0.0622$$

KOKUYO

$d_f = 29$

critical value:

$t_{0.05,29} = \ldots$

Ex 4:

$H_0: \mu_1 - \mu_2 = 0$ $\quad s_1 = 0.576$

$H_1: \mu_1 - \mu_2 \neq 0$ $\quad s_2 = 0.776$

$\alpha = 0.05 \rightarrow \alpha/2 = 0.025$ $\quad n_1 = 20, n_2 = 17$

$\bar{x}_1 = 3.4, \bar{x}_2 = 3.2$

$z_{0.025} = 1.96$

a) $z_0 = \dfrac{3.4 - 3.2}{\sqrt{\dfrac{0.576^2}{20} + \dfrac{0.776^2}{17}}} = 0.877 \approx 0.88$

p value $= 2P(Z > 0.88)$

$= 2(1 - P(Z < 0.88))$

$= 2 \cdot (1 - 0.81) = 0.38$

b) $E = 1.96 \cdot \sqrt{\dfrac{0.576^2}{20} + \dfrac{0.776^2}{17}}$

$= 0.374$

Upper: $(\bar{x}_1 - \bar{x}_2) + E = 0.574$

Table for $p_1 - p_2$:

$$\hat{p}_1 = \frac{x_1}{n_1}$$

$$\hat{p}_2 = \frac{x_2}{n_2}$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$$CI : ((\hat{p}_1 - \hat{p}_2)) \pm E$$

$$E = z_{\alpha/2} \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$H_0 = p_1 - p_2 = 0$$

Test Statistic

$$z_0 = \frac{(\hat{p}_1 - \hat{p}_2) - 0}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Ex 5:

$\mu_1 = 7.6$, $\mu_2 = 6.9$, $\sigma_1 = 1.9$, $\sigma_2 = 1.7$, ~~$\alpha = 0.0$~~

$n_1 = n_2 = 50$, $z_{0.01} = 2.33$

$H_0 = \mu_1 - \mu_2 = 0$ → two tails

$H_1 = \mu_1 - \mu_2 \neq 0$

$\alpha = 0.02 \rightarrow z_{0.01} = 2.33$

p-value $= 2P(Z > |z_0|)$

$$z_0 = \frac{\mu_1 - \mu_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

$= 2.247$

→ $z_0 \in$ [ ] → fail to reject claim

7.

$\overline{x_1} = 12.0$, $\overline{x_2} = 16.3$

$s_1 = 4.2$, $s_2 = 4.9$

$n_1 = 19$, $n_2 = 17$

$\alpha = 0.05$

$H_0 = \mu_1 - \mu_2 \geq 0$

$H_1 = \mu_1 - \mu_2 < 0$ → ~~left~~ left

→ $z_\alpha = z_{0.05,29} = 170$

$t_0 = \dfrac{\overline{x_1} - \overline{x_2}}{}$

$1 < \dfrac{s_2^2}{s_1^2} < 2$ → $\sigma_1^2 = \sigma_2^2$

$s_p = \dfrac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} \approx 18.59$

$t_0 = \dfrac{\overline{x_1} - \overline{x_2}}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}} = -2.185$

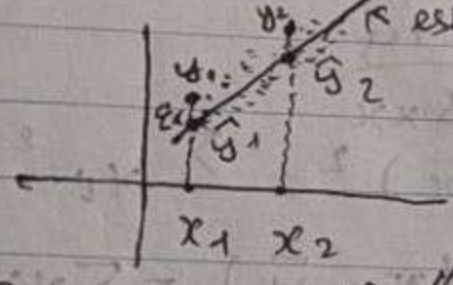
$t_{0.05,29} = 1.7$

→ $t_0 \in$ [ ]

CHAPTER 11: Simple linear regression and Correlation

* Regression line:
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

* Simple linear regression model
$$y = \beta_0 + \beta_1 x + \varepsilon$$



$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$

$\varepsilon_i = y_i - \hat{y}_i$: predicted value
- random error
- residual error

observated at $x = x_i$

* β_0: y-intercept of population regression line
β_1: slope of population regression-line

estimated y-intercept $\hat{\beta}_0 \rightsquigarrow \beta_0$
estimated slope $\hat{\beta}_1 \rightsquigarrow \beta_1$
$\hat{\sigma}^2 \rightsquigarrow \sigma^2$

$$SS_E = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$\beta_0, \beta_1 \rightsquigarrow$ min

Def: The fitted or estimated regression line for data X, Y

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}, \quad \bar{x} = \frac{\Sigma x_i}{n}, \quad \bar{y} = \frac{\Sigma y_i}{n}$$

$$\hat{\beta}_1 = \frac{\Sigma x_i y_i - \frac{\Sigma x_i \Sigma y_i}{n}}{\Sigma x_i^2 - \frac{(\Sigma x)^2}{n}} = \frac{S_{xy}}{S_{xx}}$$

Note: $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$

Sample var of x

$$S_{xx} = \Sigma (x_i - \bar{x})^2 = \Sigma x_i^2 - \frac{(\Sigma x_i)^2}{n} = (n-1) S_x^2$$

$$S_{xy} = \Sigma (x_i - \bar{x})(y_i - \bar{y}) = \Sigma x_i y_i - \frac{(\Sigma x_i)(\Sigma y_i)}{n}$$

* Standard error of estimate

- Total variation (total sum of square)

$$SS_T = S_{yy} = \Sigma (y_i - \bar{y})^2$$

- Error sum of square

$$SS_E = \Sigma (y_i - \hat{y}_i)^2$$ (unexplained variation)

- Regression sum of squares:

$$SS_R = \Sigma (\hat{y}_i - \bar{y})^2 = \beta_1 S_{xy}$$

$$SS_T = SS_R + SS_E$$

Thứ
Ngày . . No.

Def : unbiased estimator of standard error of estimate σ

$$\hat{\sigma} = \sqrt{\frac{SS_E}{n-2}}$$

* Coefficient of determination

$$R^2 = \frac{SS_R}{SS_T}$$

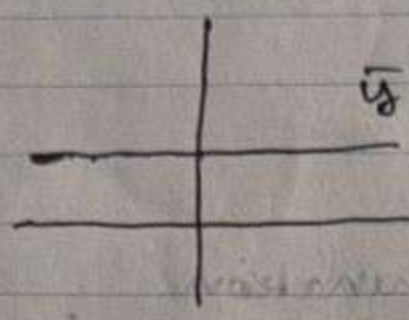
- % variation of Y is explained by % variation of X
- %>>> => cor relation: strong (correlation $\in R^2$)

Note : $0 \leq R^2 \leq 1$

+ $R^2 = 1 \Leftrightarrow SS_R = SS_T$
$\Leftrightarrow SS_E = 0 \Leftrightarrow \sum (y_i - \hat{y})^2 = 0$
$\Leftrightarrow y_i = \hat{y}$

$+ R^2 = 0 \Leftrightarrow SS_R = 0 \Leftrightarrow \sum (\hat{y}_i - \bar{y})^2 = 0$

$\Leftrightarrow \hat{y}_i = \bar{y}$



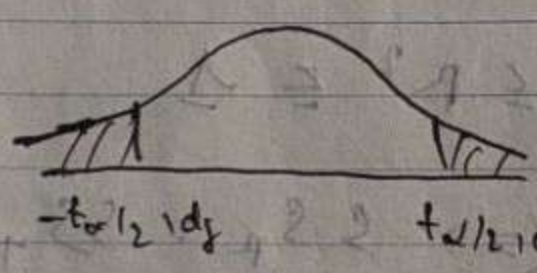
→ ∞ no correlation
→ don't use $\hat{y}_i$ to predict y
→ use $\bar{y}$ for predict $\hat{y}_i$

1. Test of hypothesis about the slope β_1

$H_0 = \beta_1 = \beta_{1,0}$ (Excel : $H_0 = \beta_1 = 0$

$H_1 = \beta_1 \neq \beta_{1,0}$ $H_1 : \beta_1 \neq 0$)

$df = n - 2$



Test statistic

$$t_0 = \frac{\hat{\beta}_1 - \beta_{1,0}}{se(\hat{\beta}_1)}$$

estimated standard error of slope

$$se(\hat{\beta}_1) = \sqrt{\frac{\sigma^2}{S_{xx}}}$$

excel : Test

$H_0 : \beta_1 = 0$ (no correlation)

$H_1 : \beta_1 \neq 0$ (∃ significant correlation)

Test $H_0 : \beta_1 = 0$
$H_1 \quad \beta_1 \neq 0$

• Reject H_0: ∃ a significant correlation between x and y
⇒ the best ~~pre-dicted~~ predicted value of y at $x = x_i$ is $\hat{y} = \hat{\beta_0} + \hat{\beta_1} x$
• Fail to reject H_0: no correlation
⇒ the best predicted value of y at $x = x_i$ is

$$\bar{y} = \frac{\Sigma y_i}{x}$$

2. Test of hypothesis about the intercept coefficient β_0.
- ý như slope β_1
- Khác: Tính chất test statistic

$$t_0 = \frac{\beta_0 - \beta_{0,0}}{se(\beta_0)}$$

↑ estimated standard error of intercept

$$se(\hat{\beta_0}) = \sqrt{\hat{\sigma}^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}} \right)}$$

* Continue:
④ Sample coefficient of correlation

$$R = \frac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$$

$$S_{xy} = \Sigma x_i y_i - n \bar{x} \bar{y}$$
$$S_{xx} = \Sigma x_i^2 - n \bar{x}^2$$
$$S_{yy} = \Sigma y_i^2 - n \bar{y}^2$$

- với ý: $S_{xx} = (n-1) s_x^2$
$S_{yy} = (n-1) s_y^2$

KOKUYO

Thứ
Ngày No.

11)

$\hat{y} = 10725.802 + 69.960x$

$\hat{y}(30) = 10725.802 + 69.960 \cdot 30$

Revision :

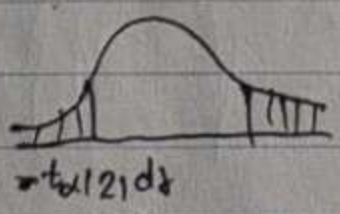
$R = \dfrac{S_{xy}}{\sqrt{S_{xx} S_{yy}}}$ correlation coefficient

Remark : R, $\hat{B_1}$ same sign

$\hat{\rho}$: $R \rightsquigarrow \rho$ unbiased estimator

sample coefficient — Population correlation coefficient

- Test claim



→ test statistic

$$t_0 = \frac{R - \rho}{\sqrt{\frac{1-R^2}{n-2}}}, \quad df = n-2$$

Ex7: Test claim $\rho = 0$, $R = 0.833$, $n = 12$, $\alpha = 0.05$

$H_0 : \rho = 0$

$H_1 : \rho \neq 0$

$$t_0 = \frac{0.833 - 0}{\sqrt{\frac{1-0.833^2}{12-2}}} = 4.761$$

$t_{\alpha/2, df} = t_{0.025, 10} = \pm 2.23$

→ $t_0 \in$ ▨ → reject H_0

→ There are

Ex??:

x	5	10	16	6	10	9
y	69	16	69	80	59	87

$\alpha = 0.05$, $t_{0.025, 4} = 2.78$,

? Test claim there are regressed in linear correlation between x and y?

$\Rightarrow \rho \neq 0$

$H_0 = \rho = 0$ (no relation)

$H_1 = \rho \neq 0$

$r = 0.2242$

$$t_0 = \frac{0.2242 - 0}{\sqrt{\frac{1 - 0.2242^2}{4}}} = 0.46$$

$\pm t_{0.025, 4} = \pm 2.78$

→ fail to reject H_0

→ don't use $\bar{y}$ to predict y_i value at x_i

→ use $\bar{y}$ to predict

Ex??

x	0	3	4	5	12
y	1	2	6	9	12

$\alpha = 0.05$

a) Test claim there are significant linear ~~regression~~ corelation between x and y?

b) Find the best predicted value of y at x=10

KOKUYO