MAS

Chapter 4:

Continuous random vraiables and Probability distibution

1.) Continuous random variable.

- Definition: A continuous random variables is a random variable whose possible values includes in an interval of real numbers.
  một khoảng

E.g:) Height: 180 cm - 190 cm

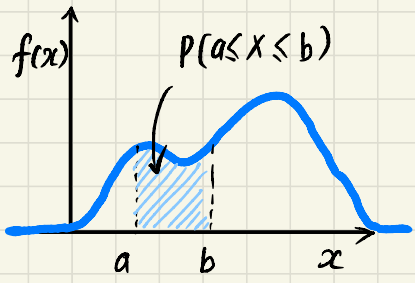
) Weight: 45.6 -> 70.2 kg

2.) Probability density function (pdf)

(Hàm mật độ xác suất) : Denote as: $f(x)$

- Definition: The propability density function (pdf) of continuous random variables X is a function such that:

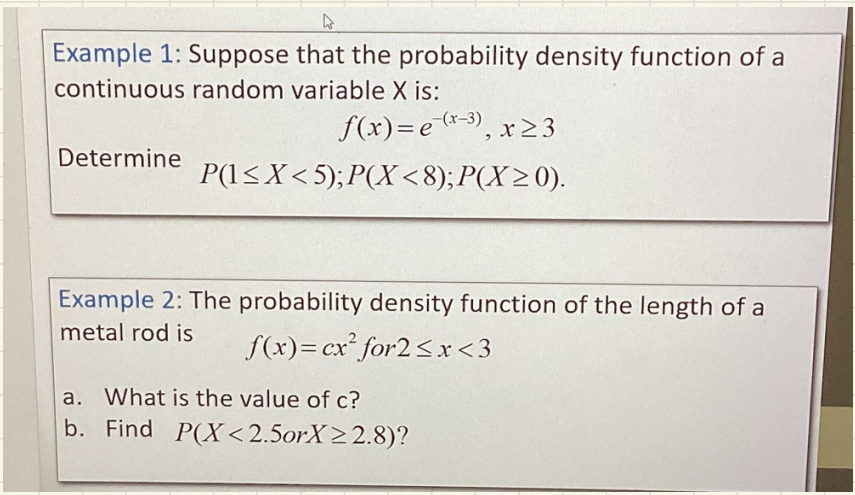
1. $f(x) \geq 0$, $\forall x$
2. $\int_{-\infty}^{+\infty} f(x)\,dx = 1$
3. $P(a \leq X \leq b) = \int_{a}^{b} f(x)\,dx$



So P is an area from a to b ← Total S = 1 (area)

Remark: Property : X is a continuous random variable (tính chất)

$$P(x_1 \leq X \leq x_2) = P(x_1 \leq X < x_2) = P(x_1 < X \leq x_2) = P(x_1 < X < x_2)$$



Example 1: Suppose that the probability density function of a continuous random variable X is:

$$f(x) = e^{-(x-3)},\ x \geq 3$$

Determine $P(1 \leq X < 5); P(X < 8); P(X \geq 0).$

Example 2: The probability density function of the length of a metal rod is $f(x) = cx^2\ for 2 \leq x < 3$

a. What is the value of c?
b. Find $P(X < 2.5 or X \geq 2.8)$?

(img 1)

Eg1: pdf: $f(x) = \begin{cases} e^{-(x-3)}, & \text{if } x \geqslant 3 \\ 0, & \text{else where} \end{cases}$

a.) $P(1 \leqslant X < 5) = \int_a^b f(x)\,dx = \int_1^5 f(x)\,dx$

$= \int_3^5 e^{-(x-3)}\,dx = 0.865$

b.) $P(X < 8) = \int_3^8 e^{-(x-3)}\,dx = 0.993262$

c.) C1: $P(X \geqslant 0) = \int_3^{+\infty} e^{-(x-3)} = \ldots$

C_2: $P(X \geq 0) = 1 - P(X < 0) = 1 - 0 = 1$

3.) Cumulative distribution function (cdf)
(Hàm phân bố tích luỹ)

- Definition: The cumulative distribution function (cdf) of a continuous random variable X is

$$F(x) := \int_{-\infty}^{x} f(t)\,dt \; (= P(X \leq x) = P(X < x))$$

(for $-\infty < x < +\infty$)

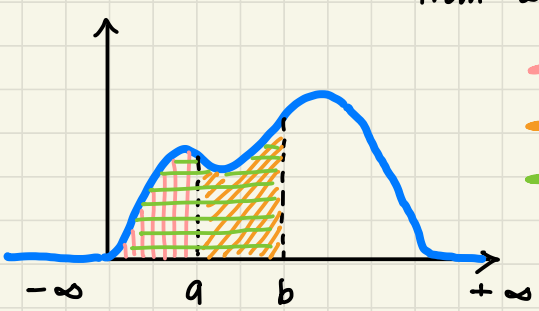
- Remark: If X is continuous randomvariable with cumulative distribution function $F(x)$ then we can use:

$$P(a < X < b) = F(b) - F(a)$$

This is the area from a to b

The area from $-\infty$ to b

The area from $-\infty$ to a



- Remark(2):

⇒ $F'(x) = f(x)$

⇒ $0 \leq F(x) \leq 1$, $\forall x$

(F is non-decreasing : $x < y \Rightarrow F(x) < F(y)$)

- $P(X = a) = 0 \quad (\forall a)$
- $P(a < X < b) = \int_a^b f(x)\,dx$
- $P(X > a) = 1 - P(X < a)$

Cumulative distribution function (cdf)

Example 1: Suppose the cumulative distribution function of the random variable X is

$$F(x) = \begin{cases} 0, & if\ x < 1 \\ 0.5x - 0.5, & if\ 1 \le x < 3 \\ 1, & if\ x \ge 3 \end{cases}$$

Find $P(X < 2.8); P(0 < X < 1.5)$?

(img 2)

Example 2: Suppose the probability density function of the random variable X is

$$f(x) = e^{-(x-3)},\ x \ge 3$$

Find the cumulative distribution function of X.

Eg 1:

$$\text{cdf: } F(x) = \begin{cases} 0, & \text{if } x < 1 \\ 0.5x - 0.5, & \text{if } 1 \le x < 3 \\ 1, & \text{if } x \ge 3 \end{cases}$$

a.) $P(X < 2.8) = F(2.8)$
$= 0.5 \times (2.8) - 0.5$
$= 0.9$

b.) $P(0 < X < 1.5) = F(1.5) - F(0)$
$= 0.25 - 0 = 0.25$

Eg2: pdf: $f(x) = c x^2 \quad (2 \leq x \leq 3)$

(img 1)

a.) $\int_{-\infty}^{+\infty} f(x)\,dx = 1$

$\Rightarrow \int_2^3 f(x)\,dx = 1 \Leftrightarrow c\int_2^3 x^2\,dx = 1$

$\Leftrightarrow c \cdot \frac{19}{3} = 1 \Rightarrow c = \frac{3}{19}$

b.) $P(X < 2.5 \text{ or } X \geq 2.8)$

$= P(X < 2.5) + P(X \geq 2.8)$

$= \int_2^{2.5} \frac{3}{19} x^2 dx + \int_{2.8}^3 \frac{3}{19} x^2 dx$

Eg2 (img2)

C_1: pdf: $f(x) = e^{-(x-3)}, \quad x \geq 3$

We have $F'(x) = f(x) = e^{-(x-3)}$

$\Rightarrow F(x) = c + [-e^{-(x-3)}]$

$\Rightarrow$ check the answer

C_2: $F(x) = \int_{-\infty}^x f(t)\,dt = \int_{-\infty}^x e^{-(t-3)}\,dt$

...

4.) Mean and Variance of a continuous random variable

- **Definition:** Suppose X is a continuous random variable with probability density function $f(x)$.

=> The Mean or expected value of X is defined by

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

=> The Variance of X is defined by

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

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

=> The standard deviation of X is:

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

Example 1: Assume that X is a continuous random variable with the following probability density function

$$f(x) = \frac{x^2}{18}, \text{ for } -3 < x < 3$$

Determine the mean and variance of X.

Example 2: The cumulative distribution function of the random variable X is

$$F(x) = \begin{cases} 0, & \text{if } x < 1 \\ 0.5x - 0.5, & \text{if } 1 \le x < 3 \\ 1, & \text{if } x \ge 3 \end{cases}$$

Find the standard deviation of X

Eg1: pdf: $f(x) = \frac{x^2}{18} \quad (-3 < x < 3)$

$\Rightarrow \mu = E(X) = \int_{-3}^{3} x \cdot \frac{x^2}{18} dx = 0$

$\Rightarrow \sigma^2 = V(X) = \int_{-3}^{3} x^2 \cdot \frac{x^2}{18} dx - \mu^2 = \frac{27}{5}$

Eg2: cdf: $F(X) := \begin{cases} 0, & \text{if } x < 1 \\ 0.5x - 0.5, & \text{if } 1 \leq x < 3 \\ 1, & \text{if } x \geq 3 \end{cases}$

$f(x) = F'(X) = 0.5 \quad (1 \leq x < 3)$

$\Rightarrow \mu = \int_{1}^{3} x \times 0.5 \, dx = 2$

$\Rightarrow \sigma^2 = V(X) = \int_{1}^{3} x^2 \times 0.5 dx - \mu^2 = \frac{1}{3}$

$\Rightarrow \sigma = \sqrt{V(X)} = \underline{\sqrt{3}}$

5.) Continuous uniform distribution

- Definition: Suppose X has a continuous uniform distribution over the interval [a, b] if

- pdf: $f(x) = \begin{cases} \frac{1}{b-a} & , \text{if } a \le x \le b \\ 0 & , \text{otherwise} \end{cases}$

- Mean & Variance:

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

$= \int_a^b \frac{1}{b-a} \cdot x\,dx = \frac{a+b}{2}$

+ $\sigma^2 = V(X) = E(X)^2 - (E(X))^2$

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

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

Example 1: Suppose X has a continuous uniform distribution over the interval [1;10].

- Determine the mean, variance and standard deviation of X.
- Find P(X<6.5).
- Determine the cumulative distribution function of X.

Example 2: Suppose X has a continuous uniform distribution over [5;15]. What is the mean and variance of Y=8X?

Eg 1: X has continuous uniform distribution on [1; 10]

$\Rightarrow [a, b] = [1, 10] \Rightarrow \begin{cases} a = 1 \\ b = 10 \end{cases}$

a.)

- Mean $= \mu = \frac{a+b}{2} = 5.5$
- Variance $= \sigma^2 = \frac{(b-a)^2}{12} = 6.75$
- Standard deviation $= \sigma = \sqrt{\sigma^2} = 2.6$

b.)

$$\text{pdf: } f(x) = \begin{cases} \frac{1}{b-a} = \frac{1}{9}, & \text{if } x \in [1, 10] \\ 0, & \text{if } x \notin [1, 10] \end{cases}$$

$$\Rightarrow P(X < 6.5) = \int_1^{6.5} f(x)\,dx = 0.61$$

c.)

$$\text{cdf: } F(X) := \begin{cases} 0, & \text{if } x < 1 \\ \frac{x-1}{9}, & \text{if } 1 \leq x < 10 \\ 1, & \text{if } x \geq 10 \end{cases}$$

Eg2: $Y = 8X$

- $E(Y) = E(8X) = 8E(X) = 8 \cdot \frac{(a+b)}{2}$
- $V(Y) = V(8X) = 8^2 V(X)$
- $\sigma(Y) = \sqrt{V(Y)}$

6.) Normal distribution (Phân phối chuẩn)

- Definition: X has a normal random variable with parameters μ and σ^2 if

(Notation: $X = N(\mu, \sigma^2)$)

- pdf: $f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{-(x-\mu)^2}{2\sigma^2}}, \quad -\infty < x < +\infty$

When $-\infty < \mu < +\infty$ and $\sigma > 0$

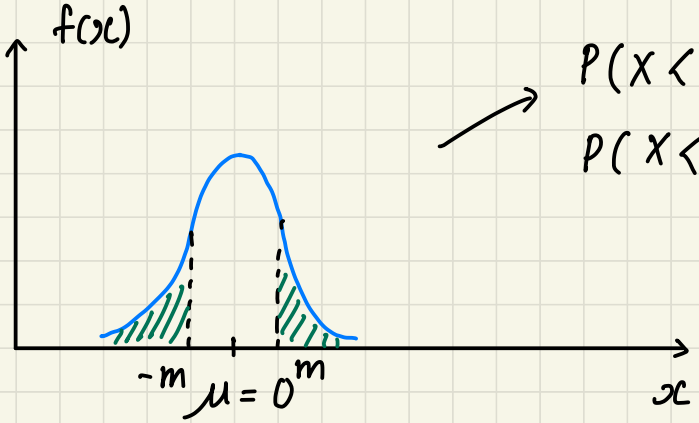
- Mean: $\mu = E(X)$
- Variance: $V(X) = \sigma^2$
- Standard deviation: $\sigma = \sqrt{V(X)}$

7.) Standard normal distribution (Phân phối chuẩn tắc)

- Notation: $Z \underset{(\sim)}{=} N(0,1)$

$\Rightarrow \mu = 0$ and $\sigma^2 = 1 \Rightarrow \sigma = 1$

- pdf: $f(x) = \frac{1}{\sqrt{2\pi}} \cdot e^{\frac{-x^2}{2}}, \quad -\infty < x < +\infty$
- cdf: $\Phi(z) = \int_{-\infty}^{z} \frac{e^{\frac{-x^2}{2}}}{\sqrt{2\pi}} dx$



$P(X < -m) = P(X > m)$

$P(X < m) = 1 - P(X > m)$

8.) Normal Standardizing (Chuẩn tắc hóa)

- Definition: If $X = N(\mu, \sigma^2)$ then

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

is a standard normal random variable $N(0,1)$

$$\left(X \sim N(\mu, \sigma^2) \Rightarrow Z = \frac{X-\mu}{\sigma} \sim N(0,1) \right)$$

$$\therefore P(X \lessgtr x) = P\left(\frac{X-\mu}{\sigma} \lessgtr \frac{x-\mu}{\sigma}\right)$$

$$= P\left(Z \lessgtr \frac{x-\mu}{\sigma}\right)$$

$$\therefore P(a < X < b) = P(X < b) - P(X < a)$$

Eg: The weekly salaries are normally distributed with a means of \$490 and standard deviation of \$45.

Find P (a teacher earns $> \$525$)

Solution: $\mu = 490$, $\sigma = 45$

$$P(X > 525) = 1 - P(X < 525)$$

$$= 1 - P\left(Z < \frac{525 - 490}{45}\right)$$

$$= 1 - P(Z < 7/9)$$

$$= 0.2183$$

Example 1: A machine pours beer into 16 oz. bottles. Experience has shown that the number of ounces poured is normally distributed with a standard deviation of 1.3 ounces. Find the probabilities that the amount of beer the machine will pour into the next bottle will be more than 16.25 ounces.

Example 2: The tread life of a particular brand of tire is a random variable best described by a normal distribution with a mean of 65,000 miles and a standard deviation of 1,500 miles. What warranty should the company use if they want 95% of the tires to outlast the warranty?

Eg 1: $\mu = 16$; $\sigma = 1.3 \Rightarrow \sigma^2 = 1.69$

$\Rightarrow X \sim N(16, 1.69)$

$$P(X > 16.25) = 1 - P(X \leq 16.25)$$

$$= 1 - P\left(Z < \frac{5}{26}\right) = 0.42375$$

Eg 2: $\mu = 65,000$; $\sigma = 1,500$

X: the tread life of a brand of tire

x: the warranty

$P(X > x) = 0.95$

$\Rightarrow \quad P(X < x) = 0.05$

$\hookrightarrow \quad P\left(Z < \dfrac{x - 65,000}{1500}\right) = 0.05$

$\Rightarrow \quad \dfrac{x - 65,000}{1500} = -1.645$

$\Rightarrow x = 62,532.5$ (miles)

9.) Approximation using Normal Distribution

a. Normal approximation to the Binomial distribution
(Xấp xỉ phân phối nhị thức bởi phân phối)

Definition: If X has a Binomial distribution $B(n,p)$ then random variable:

$$Z = \frac{X - n.p}{\sqrt{n.p.(1-p)}} \qquad \begin{cases} E(X) = n.p \\ \sigma(X) = \sqrt{n.p(1-p)} \end{cases}$$

$$\approx N(0,1)$$

is approximately a standard random variable $(0,1)$

We have:

$$P(X \leq x) \approx P\left(Z \leq \frac{x + 0.5 - n.p}{\sqrt{n.p.(1-p)}}\right)$$

$$P(X \geq x) \approx P\left(Z \geq \frac{x - 0.5 - n.p}{\sqrt{n.p(1-p)}}\right)$$

Remark: The approximation is good for $n.p > 5$ and $n(1-p) > 5$

Example: The manufacturing of semiconductor chips produces 3% defective chips. Assume the chips are independent and that a lot contains 800 chips. Approximate the probability that more than 30 chips are defective.

Hint: X~B(n = 800; p = 0.03). Use Normal approximation.

Remark: Can use BINOM.DIST in Excel to find actual value.

Eg: $X \sim B(n = 800; p = 0.03)$

$$P(X > 30) = 1 - P(X \leq 30)$$

$$\approx 1 - P\left(Z \leq \frac{30 \overset{+0.5}{-} 800 \times 0.03}{\sqrt{800 \times 0.03 \times 0.97}}\right)$$

$$= 1 - P(Z \leq 1.347)$$

$$= 0.9111$$

b.) Normal Approximation to the Poisson Distribution

* λ : mean number of events in a unit interval
  X : number of event in a unit interval

Definition: If X has Possion distribution $P(\lambda)$ then random variable:

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

$$\begin{cases} E(X) = \mu = \lambda \\ \sigma^2 = V(X) = \lambda \end{cases}$$

is approximately a standard random variable $N(0,1)$

We have:

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

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

Remark: The approximation is good for $\lambda > 5$

Example: The number of customers that arrive at a fast-food business during a one-hour period is known to be Poisson distributed with a mean equal to 9.6. What is the probability that more than 10 customers will arrive in a one-hour period?

Hint: X~P(9.6). Use Normal approximation

Remark: Can use POISSON.DIST in Excel to find actual value.

Eg: $\mu = 9.6 = \lambda$

$$P(X > 10) = 1 - P(X \leq 10)$$

$$\approx 1 - P\left(Z \leq \frac{10 + 0.5 - 9.6}{\sqrt{9.6}}\right)$$

$$= 0.386$$

10.) Exponential distribution (phân phối mũ)

- λ : mean number of event in a unit interval
- X : the distance between two consecutive event

Definition: The random variable X that equals the distance between two consecutive events with mean number of event $\lambda > 0$ per unit interval is an exponential random variable with parameter λ. The probability density function of X is:

pdf: $$f(x) = \lambda e^{-\lambda x}, \quad x \geq 0$$

cdf: $$F(x) = 1 - e^{-\lambda x}, \quad x \geq 0$$

Remark: If random variable X has exponential distribution with parameter λ then $\mu = E(X) = \frac{1}{\lambda}$; $\sigma^2 = V(X) = \frac{1}{\lambda^2}$

Example 1: The time between customer arrivals at a furniture store has an approximate exponential distribution with mean of 9 minutes. If a customer just arrived, find the probability that the next customer will not arrive for at least 15 minutes.

Example 2: The time between patients arriving at an outpatient clinic follows an exponential distribution at a rate of 15 patients per hour. What is the probability that a randomly chosen arriving interval will not exceed 6 minutes?

Eg1: X = the time between customer arrivals

$\Rightarrow X \sim \exp(\lambda)$

We have mean of $X = 9$ minute $\Rightarrow \mu = \frac{1}{\lambda} = 9$

$\Rightarrow \lambda = \frac{1}{9}$ (arrivals / minute)

$$P(X > 15) = 1 - P(X \leq 15)$$

$$= 1 - \int_0^{15} \frac{1}{9} . e^{-\frac{1}{9}x} dx$$

$$= 0.189$$

Eg2: X : the time between arrivals (in hours)

$\lambda = 15$ (patient / hour)

$$P(X \leq 0.1) = \int_0^{0.1} 15 . e^{-15x} dx$$

$$= 0.777$$

Chapter 6:

Data description

Introduction to statistic

Statistic

Descriptive Statistic
(Thống kê mô tả)

Use numerical summaries and visual displays to discribe data
(Chapter 6)

Inferential Statistic
(Thống kê suy luận)

Use information from samples (data) to estimate for population
(Chapter 8, 9, 10, 11)

1.) Numerical summarise of data

a.) Sample mean (Trung bình mẫu)

Definition: If the n observations in a sample are denoted by $x_1, x_2, \ldots, x_n$, the sample mean is

$$\overline{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

(n is the size of sample)

Example: Let's consider the weight of the eight observations collected from the prototype engine connectors: 12.6, 12.9, 13.4, 12.3, 13.6, 13.5, 12.6 and 13.1

Find the sample mean.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{12.6 + 12.9 + \ldots + 13.1}{8} = 13.0$$

b.) Sample median (Trung vị mẫu)

Definition: - The value that lies in the middle of the data when the data set is ordered.

- Measures the center of an ordered data set by dividing it into two equal part.

- If the data set has an:
  (a) even number of entries: median is the avarage of the two middle data entries.

  (b) odd number of entries: median is the middle data entries.

Example: The prices (in dollars) for a sample of round-trip flights from Chicago, Illinois to Cancun, Mexico are listed. Find the median of the flight prices.

872 432 397 427 388 782 397

Eg: Order the data:

388, 397, 397, 427, 432, 782, 872

The sample median is 427.

c.) Sample mode

- The data entry that occurs with the greast frequency
- If no entry is repeated, the data set has no mode
- If two entries occur with the same greastest frequency, each entry is a mode (bimodal)

d.) Sample variance and sample standard deviation

- If $x_1, x_2, \ldots, x_n$ is a sample of n observations, the sample variance is

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

$$= \frac{\sum_{i=1}^{n} x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} x_i\right)^2}{n-1}$$

(s = sample standard deviation)

Example: Let's consider the weight of the eight observations collected from the prototype engine connectors: 12, 13, 9, 12, 10 and 12.

Find the sample standard deviation.

e.) Sample range (khoảng biến thiên)

- Sample range $= R = \max\{x_i\} - \min\{x_i\}$

2.) Stem and leaf diagram (Sơ đồ thân và lá)

Stem and leaf diagram

A stem-and-leaf diagram is a good way to obtain an informative visual display of a data set where each number xi consists of at least two digits. To construct a stem-and-leaf diagram, use the following steps:

- Divide each number x_i into two parts: a stem, consisting of one or more of the leading digits, and a leaf, consisting of the remaining digit.
- List the stem values in a vertical column.
- Record the leaf for each observation beside its stem.
- Write the units for stems and leaves on the display.

* Remark: - Có bao nhiêu số thì có bấy nhiêu lá

⇒ the number of stems $\leq n$

⇒ the number of leaves $= n$

* Note: If leaf unit = 0.1

=> 316 represent 3.6

unit = 100 (6 x 100 = 600) ← put 3 in here

=> 316 represent 3600

Stem and leaf diagram

Example: The listening scores of 12 students in a TOEIC test are listed below

55 115 225 240 330 335 385 400 405 405 495 495

The stem and leaf diagram:

Stem	Leaves
5	5
11	5
22	5
24	0
33	0 5
38	5
40	0 5 5
49	5 5

3.) Box-plots (Biểu đồ hộp)

phần tử

Three quartiles

An ordered set of data is divided into four equal parts, the division points are called quartiles:

- The first quartile, q_1 or Q_1: is a value that has approximately 25% of the observations below.
- The sample median or second quartile, q_2 or Q_2, has approximately 50% of the observations below its value.
- The third quartile, q_3 or Q_3, has approximately 75% of the observations below its value.
- The interquartile range, $IQR = Q_3 - Q_1$ (chênh lệch tứ phân vị)

Example: Use the given sample data to find the sample quartiles, the sample mode and the IQR.

55, 52, 52, 52, 49, 74, 67, 55.

* Way to determine Q_1, Q_2, Q_3:

1.) Find Q_2 = sample median (Q_2 là trung vị mẫu)

2.) Find Q_1 = sample median of the Q_2's left part

3.) Find Q_3 = _____ right part

4.) Compute IQR (interquartile range = chênh lệch tứ phân vị)

$$IQR = Q_3 - Q_1$$

5.) Compute $\begin{cases} Q_1 - 1.5 * IQR & (1) \\ Q_3 + 1.5 * IQR & (2) \end{cases}$

6.) Outlier (phần tử ngoại lai): an entry is called outlier if it smaller than (1) or greater than (2)

7.) Minimum: is the smallest entry (not an outlier)

Maximum: is the biggest entry (not an outlier)

Eg: Data set: 49, 52, 52, 52 | 55, 55, 67, 71

Q_2 = sample median = $\frac{52+55}{2} = 53.5$

$Q_1 = \frac{52+52}{2} = 52$

$Q_3 = \frac{55+67}{2} = 61$

- $IQR = Q_3 - Q_1 = 61 - 52 = 9$
- $Q_1 - 1.5 * IQR = 38.5$
- $Q_3 - 1.5 * IQR = 74.5$

=> does not have outlier

=> $\begin{cases} \text{Minimum} = 49 \\ \text{Maximum} = 71 \end{cases}$

4.) Frequency distribution (Phân bố tần suất)

Frequency Distribution

Construction of frequency distribution: divide the range of the data into intervals (called class intervals, cells, or bins). The bins should be of equal width.

Example:
Data = Grades = {2.4, 4.4, 4.6, 5.0, 5.0, 5.8, 6.0, 7.4, 8.2, 9.0}

- Divide grade ranges into 5 bins: 0 - 2, 2 - 4, 4 - 6, 6 - 8, 8 - 10.
- Count the number of data values in each bin:

Bin	Frequency
0 - 2	0
2 - 4	1
4 - 6	6
6 - 8	1
8 - 10	2

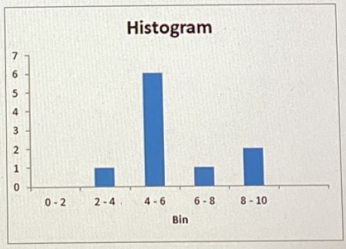
* Remark:

- frequency: tần số (n_i)
- relative frequency: tần suất ($f_i = \frac{h_i}{n}$)

Histogram

The **histogram** is a visual display of the frequency distribution.

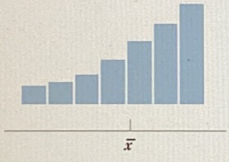
- Label the bin (class interval) boundaries on a horizontal scale.
- Mark and label the vertical scale with the frequencies or the relative frequencies.
- Above each bin, draw a rectangle where height is equal to the Frequency (or relative frequency) corresponding to that bin.



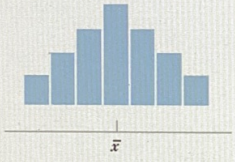
Histogram

Remark:

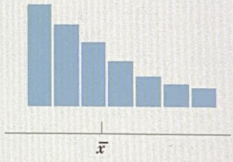
1. Histograms are very useful to explore the distribution of data.



Negative or left skew

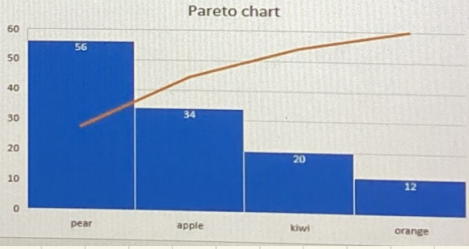


Normal



Positive or right skew

2. Pareto chart: (frequencies are ordered decreasingly)



Times sequence plots

A time series plot is a graph in which the vertical axis denotes the observed value of the variable (say, x) and the horizontal axis denotes the time (which could be minutes, days, years, etc.)

Years	Annual salary
2005	20614
2006	21175
2007	24766
2008	28100
2009	30189
2010	24618
2011	22006
2012	30912
2013	32523
2014	35285

Chapter 7:

Sampling distribution and Point estimates of parameter

1. Point estimates of parameters.
(Các ước lg điểm của các tham số.)

Parameters of Population (Các tr số của tổng thể)	Statistics of a sample (Các thống kê của mẫu)
μ: population mean	$\bar{x}$: sample mean
σ^2: variance	s^2: sample variance
σ: standard deviation	s: " standard deviation
p: population proportion (tỉ lệ của tổng thể)	$\hat{p}$: sample proportion (tỷ lệ mẫu)

Definition: A point estimate for a parameter is a single value computed from a sample

$\Rightarrow$ The best point estimate for μ is the $\bar{x}$

σ^2 is the s^2

σ is the s

Example 1: In a sample of 73 products, there are 7 defective products. So a point estimate for proportion p of all defective products is:

$$\hat{p} = \frac{x}{n} = \frac{7}{73} = 9.6\%$$

Example 2: Market researchers use the number of sentences per advertisement as a measure of readability for magazine advertisements. The following represents a random sample of the number of sentences found in 15 advertisements.

9 20 18 16 9 9 11 13 22 16 5 18 6 6 5

a. Find a point estimate of the population mean, μ.

b. Find a point estimate of the population standard deviation, σ.

Hint:

$$\bar{x} = \frac{\sum x_i}{n} = 12.2; s = \frac{\sum (x_i - \bar{x})^2}{n-1} = 5.77$$

Eg 1: A point estimate for proportion p of all deffective product is $\hat{p}$

$$\hat{p} = \frac{x}{n} = \frac{7}{73} = 0.096 = 9.6\%$$

Eg2: a.) A p/e of the population mean μ is $\bar{x}$

$$\bar{x} = \frac{\Sigma x_i}{n} = 12.2$$

b.) A p/e of the population standard deviation σ is s

$$s^2 = \frac{\Sigma (x_i - \bar{x})^2}{n-1} = 5.77$$

$$\Rightarrow s = \sqrt{s^2}$$

2.) Central limit theorem (Định lí giới hạn trung tâm)

Theorem 1. (CLT for 1 population)

- Consider a population with mean $= \mu$; variance $= \sigma^2$
- A random sample of size n: $X_1, X_2, \ldots, X_n$

Sample mean: $\overline{X} = \dfrac{X_1 + X_2 + \ldots + X_n}{n}$

Then: 1) The sampling distribution (phân bố mẫu) of sample mean is

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right), \text{ for } n \geq 30$$

(N: normal distribution)

$\Rightarrow$ It means that mean $= \mu$, variance of $\overline{X}$ is $\dfrac{\sigma^2}{n}$

> Example: One year, professional players salaries averaged 1.5 million with a standard deviation of 0.9 million. Suppose a sample of 100 players was taken. Find the approximate probability that the average salary of these 100 players does not exceed 1.4 million.

Eg: Population mean (ở đây là trung bình) $= \mu = 1.5$; standard deviation $= \sigma = 0.9$

Size of a sample: $n = 100$.

$$P(\overline{X} < 1.4) = P\left(\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < \frac{1.4 - \mu}{\sigma/\sqrt{n}}\right)$$

$$= P\left(Z < \frac{1.4 - 1.5}{0.9/\sqrt{100}}\right) = P(Z < -1.111) = 0.1333$$

Remark: $\Rightarrow Z = \dfrac{\overline{X} - \mu}{\delta/\sqrt{n}} \sim N(0, 1)$

Phân bố chuẩn tắc

$(n \geq 30)$

⇒ If the population already has normal distribution then for any sample size:

$$\overline{X} \sim N\left(\mu; \frac{\delta^2}{n}\right)$$

Example: An electrical firm manufactures light bulbs that have a length of life that is approximately normally distribution, with mean equal to 800 hours and a standard deviation of 40 hours. Find the probability that a random sample of 16 bulbs will have an average life of less than 775 hours.

Eg2: Suppose that population is normal with mean $= \mu = 800$, standard deviation $= \delta = 40$.

Sample size : $n = 16$.

We have :

$$P(\overline{X} < 775) = P\left(\frac{\overline{X} - \mu}{\delta/\sqrt{n}} < \frac{775 - \mu}{\delta/\sqrt{n}}\right)$$

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

$$= 0.006 \quad (\text{Don't need } "n \geq 30")$$

* Chú ý:

Chương 4: $P(X < b) = P\left(\dfrac{X - \mu}{\delta} < \dfrac{b - \mu}{\delta}\right)$

X : is normal random variance, $X \sim N(\mu, \delta)$

Chương 7: $P(\bar{X} < b) = P\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < \frac{b - \mu}{\sigma/\sqrt{n}}\right)$

$$= P\left(z < \frac{b - \mu}{\sigma/\sqrt{n}}\right)$$

$\bar{X}$ is sample mean of a random sample taken from a population with mean=μ, Standard deviation=σ
n: sample size

Theorem 2: (CTL for two population)

Population 1	Population 2
mean: μ_1	mean: μ_2
Standard deviation: σ_1	" " : σ_2
Sample size: n_1	Sample size: n_2
Sample mean: $\bar{X}_1$	Sample mean: $\bar{X}_2$

We have:

$$\bar{X}_1 - \bar{X}_2 \sim N\left(\underbrace{\mu_1 - \mu_2}_{\text{mean}}, \underbrace{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}_{\text{variance}}\right)$$

(for $n_1 \geq 30$, $n_2 \geq 30$)

$$z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0,1)$$

Note: $P(\bar{X}_1 - \bar{X}_2 < b)$

$$= P\left(\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} < \frac{b - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}\right)$$

$\Rightarrow$ Các xs $P(\bar{X}_1 - \bar{X}_2 > a)$ or $P(a < \bar{X}_1 - \bar{X}_2 < b)$ được tính tương tự

Example: The television picture tubes of manufacturer A have a mean lifetime of 6.5 years and a standard deviation of 0.9 year, while those of manufacturer B have mean lifetime of 6.5 years and a standard deviation of 0.8 year. What is the probability that a random sample of 36 tubes from manufacture A will have mean lifetime that is at least 1 year more than the mean lifetime of a sample of 49 tubes from manufacturer B?

Population 1	Population 2
$\mu_1 = 6.5$	$\mu_2 = 6.5$
$\sigma_1 = 0.9$	$\sigma_2 = 0.8$
$n_1 = 36$	$n_2 = 49$

Eg: Population 1 (Manufacture A) $\mu_1 = 6.5$, $\sigma_1 = 0.9$, $n_1 = 36$

Population 2 (Manufacture B) $\mu_2 = 6.5$, $\sigma_2 = 0.8$, $n_2 = 49$

We have

$$P(\bar{X}_1 - \bar{X}_2 \geq 1) = P\left(\frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma^2/n_2}} \geq \right.$$

$$\geq \frac{1-(\mu_1-\mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1}-\frac{\sigma_2^2}{n_2}}}\Bigg)$$

$$= P\left(Z \geq \frac{1-(6.5-6.5}{\sqrt{\frac{0.9^2}{36}+\frac{0.8^2}{49}}}\right) = P(Z \geq 5.302)$$

$$= 1 - P(Z < 5.302) \approx 1-1=0$$

MAS PT2: (Ryu Yamada)

1.) $P(X = 2018) = 0$. Because $P(X = a) = 0 \ (\forall a)$ (D)

2.) $P(\bar{X} > \frac{1458}{36}) = P(\bar{X} > 40.5) = 1 - P(\bar{X} \leq 40.5)$

$= 1 - P\left(Z \leq \frac{40.5 - 40}{\frac{2}{\sqrt{36}}}\right)$

$= 1 - P(Z \leq 1.5) = 0.0668$

≈ 0.067

(A)

3.) Standard error $= \frac{\sigma}{\sqrt{n}} = \frac{30}{\sqrt{9}} = 10$ (E)

4.) Cumulative relative frequence: tần số tương đối tích lũy

$\Rightarrow$ the number of... $= [P(21-30) - P(11-20)] * 80$

$= (72.5\% - 27.5\%) * 80$

$= 36$ (members) (C)

5.) 32 numbers

$Q_2 = \frac{x_{16} + x_{17}}{2} = 114$

$Q_1 = \frac{x_8 + x_9}{2} = 108$

$Q_3 = \frac{x_{24} + x_{25}}{2} = 118$

$\Rightarrow$ (B)

6.) Mode number (the number has the biggest frequency)

$= 16$ (2 times) (B)

7.) $\mu = 2.6, \quad \sigma = 0.4$

$$P(2.2 < X < 3) = P(X < 3) - P(X < 2.2)$$

$$= P\left(Z < \frac{3 - 2.6}{0.4}\right) - P\left(Z < \frac{2.2 - 2.6}{0.4}\right)$$

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

$$= 0.68 = 68\% \quad (A)$$

8.) continuous uniform distribution $[1, 5]$

$$\Rightarrow \mu = \frac{a+b}{2} = 3$$

$$\Rightarrow \sigma^2 = \frac{(b-a)^2}{12} \Rightarrow \sigma = 1.1547.$$

(C)

9.) $P(X < 950) = 1 - P(X \geq 950)$

$$\approx 1 - P\left(Z \geq \frac{950 - 0.5 - \lambda}{\sqrt{\lambda}}\right)$$

$$= 1 - P(Z \geq 1.5969)$$

$$= 0.0548 \quad (D)$$

10.) $\lambda = 5$ uniform distribution $[\underset{a}{4.5}, \underset{b}{7.5}]$

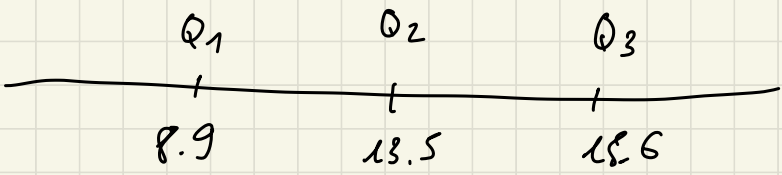
$$P(5.0 < X < 6.0) = P(X < 6.0) - P(X < 5.0)$$

$$= \int_0^6 f(x)\,dx - \int_0^5 f(x)\,dx$$

$$= \int_0^6 \frac{1}{b-a}\,dx - \int_0^5 \frac{1}{b-a}\,dx = \frac{1}{3}$$

(C)

13.) $\bar{X} = \frac{\text{Sum}}{1500} \Rightarrow \text{Sum} = 21\,300$ in



$IQR = Q_3 - Q_1 = 6.7$

$Q_1 - 1.5 * IQR = -1.15$ (= 0)) 2 Klassen

$Q_3 + 1.5 * IQR = 25.65$)

$\Rightarrow$ (E)

11.)

8	5	8
9	1	8
10	5	5

$\Rightarrow$ 85, 88, 91, 98, 105, 105

(C)

12.) expected value = mean = $\mu = \int_{-\infty}^{+\infty} x \cdot f(x)\, dx$ $\rightarrow$ $f(x) = 4x^3$, $= 1$, $= 0$

$= 0.8$ (A)

14.) exponential distribution. $\mu = 3 \Rightarrow \lambda = \frac{1}{3}$

$P(X > a)$? time $= 0.1$

$\Rightarrow P(X \leq a) = 0.9 = \int_{-\infty}^{+\infty} f(x)\, dx$

$\Leftrightarrow \int_0^a \frac{1}{3} \cdot e^{-\frac{1}{3}x}\, dx = 0.9$ (0.999)

$\Rightarrow a \approx 6.9$ (thử đáp án) (D)

MAS PT2 : Nguyên tử

1) * (Đã làm) (D)

2.) $P(-1<X<2) = F(2) - F(-1)$

$= 1 - \left[\frac{1}{2}(-1)^3 + \frac{1}{2}\right]$

$= 1$ (C)

3.) $P(X \leq 31) \approx P\left(Z \leq \frac{31 + 0.5 - n.p}{\sqrt{n.p.(1-p)}}\right)$

$= P(Z \leq 0.327)$

$= 0.6293$ (A)

4.) * (C)

5.) ex dis $\mu = 3.4 \Rightarrow \lambda = \frac{1}{3.4}$

$P(X>5) = 1 - P(X \leq 5)$

$\approx 1 - \int_0^5 \lambda . e^{-\lambda x} dx$

$= 1 - 0.77 = 0.229790$

(D)

6.) $\text{mean} = \frac{4+14+3+16+9+8+16}{7} = 10$

(D)

7.) $F(x) = 1 - e^{-x/5} \Rightarrow f(x) = F'(x)$

$= \frac{1}{5} e^{-x/5}$

$\Rightarrow \mu = E(X) = \int_0^{+\infty} x \cdot \frac{1}{5} e^{-x/5} dx \simeq 5$

(C)

8.) $p(X < 48) = p\left(Z < \frac{48-45}{6}\right)$

$= p\left(Z < \frac{1}{2}\right)$

$= 0.6915$ (C)

9.) * (A)

10.) $\bar{X}$? $n = 15$, $\mu = 6$, $\delta = 2.5$ (= mean)

Variance $= \frac{\delta^2}{n} \Rightarrow$ (B)

11.) (B)

12.) $\mu = 1$; $\delta = 0.1$

$p(X > a) = 0.025$

$\Rightarrow p(X \leq a) = 0.975$

$\Leftrightarrow p\left(Z \leq \frac{a-1}{\frac{0.1}{\sqrt{12}}}\right) = 0.975$

(C)

$\Rightarrow a = 1.05649$

13.) $P(X < 1.5) = \int_0^{1.5} 0.5\,dx = 0.75$ (B)

14.) (A)

15.) $P(X < 0 \text{ or } X > 0.5) = 1 - P(0 \leq X \leq 0.5)$

$= 1 - \int_0^{0.5} 1.5x^2 = 0.9375$ (A)

MAS PT2: Ko phải Hưng

$$\int_{-\infty}^{+\infty} \lambda \cdot e^{-\lambda x} dx$$

1.) $\mu = 3.4 \Rightarrow P(X < 1) = \int_0^1 \frac{1}{3.4} \cdot e^{-\frac{1}{3.4}x} dx$

$\Rightarrow \lambda = \frac{1}{3.4}$

$= 0.2548$ (B)

3.) $\sigma^2 = V(x) = \int_0^2 (x - \mu)^2 f(x) dx = \frac{2}{9}$ (D)

$\mu = \int_0^2 x f(x) dx = \frac{4}{3}$

2.) (C.)

4) Mean = μ

Variance = $\frac{\sigma^2}{n}$ $(x\sigma^2)$

Stadard deviation = $\sqrt{\frac{\sigma^2}{4}} \rightarrow$ o đổi.

(D)

$n \uparrow \Rightarrow sd \downarrow$

5.) $\sigma = \sqrt{\frac{\sigma^2}{n}} = \sqrt{\frac{20}{100}} = 0.447 \simeq 0.45$ (B)

6.) $\bar{X} = \frac{64.9 + 65 + 65.5}{3} = \frac{977}{15}$

$\Rightarrow s^2 = \frac{(64.9 - \bar{X})^2 + (65 - \bar{X})^2 + (65.5 - \bar{X})^2}{n-1}$

$= 0.10333$

$\Rightarrow s = \sqrt{s^2} = 0.321$ (A)

7.) * (A)

8.) (A)

9.) $f(x) = F'(x) = \frac{1}{2} e^{x/2}$

$P(X > 1) = 1 - P(X < 1)$

$= 1 - \int_0^1 \frac{1}{2} e^{-x/2} dx$

$= 0.6065$ (C)

10.) $P(13 < X < 65) = P(X < 65) - P(X < 13)$

$= \int_0^{65} \underbrace{\frac{1}{b-a}}_{70-10} dx - \int_0^{13} \frac{1}{b-a} dx$

$= 0.86$ (C)

11.) (?)

12.) $4 + 5 + 8 + 6 + 2 = 25$ (A)

13.) (?)

14.) B

15) $P(X > a) = 75.8\%$

$\Rightarrow P(X \leq a) = 0.242$

$\Leftrightarrow P\left(Z \leq \frac{a-\mu}{\sigma}\right) = 0.242$

$\Rightarrow \frac{a-3.5}{1} = -0.7$

$\Rightarrow a = 2.8$ B

Chapter 8:

Statistical intervals for a single sample

Đ Confidence interval (CI): Khoảng tin cậy

Đ Confidence level: $100.(1-\alpha)\% = 1-\alpha$ (Độ tin cậy)

Ex: Confidence level $= 95\% = 0.95 = 1-\alpha \Rightarrow \alpha = 0.05$

Đ Critical value (a percentage point): giá trị giới hạn: z_α

$$P(Z > z_\alpha) = \alpha$$

$$\Rightarrow P(Z < z_\alpha) = 1-\alpha$$

Ex: Find $z_{0.025}$ $\Rightarrow P(Z > z_{0.025}) = \alpha$

$\Rightarrow P(Z < z_{0.025}) = 0.979$

$\Rightarrow z_{0.025} = 1.96$

Confidence interval for μ (σ is known)

Theorem (C.I. for μ if σ is known)

If $\bar{x}$ is the sample mean of a random sample of size n from a normal population with known variance σ^2, a 100(1-α)% CI on μ is given by:

$$\bar{x} - z_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Example: In a sample of 36 randomly selected women, it was found that their mean height was 65.3 inches. From previous studies, it is assumed that the standard deviation of all women heights σ = 2.5 (in). Construct a 90% confidence interval for the mean height of all women.

1. Continuous interval for μ of a normal distribution σ is known

μ : population mean (unknown) } parameters

σ : population standard deviation }

a.) A $100(1-\alpha)\%$ two-sided CI on μ is given by:

$$\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

b.) A $100(1-\alpha)\%$ upper confidence bound on μ is : (chặn trên của μ)

$$\mu \leq \bar{x} + z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}}$$

c.) A $100(1-\alpha)\%$ lower confidence bound on μ is : (chặn của μ)

$$\mu \geq \bar{x} - z_{\alpha} \cdot \frac{\sigma}{\sqrt{n}}$$

* Here: $\bar{x}$ is sample mean ; n is sample size

Ex1: Suppose population is normal with standard deviation

$\sigma = 0.010$ (σ is known)

A 99% two-sided CI on μ is :

$$\bar{x} - z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}} < \mu < \bar{x} + z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$$

$n = 10$; $\bar{x} = 1.545$

Confidence level: $99\% = 1 - \alpha \Rightarrow \alpha = 0.01$

$\Rightarrow z_{\frac{\alpha}{2}} = z_{0.005} = 2.58$

$$1.545 - 2.58 \times \frac{0.010}{\sqrt{10}} < \mu < 1.545 + 2.58 \times \frac{0.010}{\sqrt{10}}$$

$$\Leftrightarrow \quad 1.537 < \mu < 1.553$$

Answer: $\mu \in (1.537, 1.553)$

b.) A 90% CI on μ is: $\left(z_{\frac{\alpha}{2}} = z_{0.05} = 1.645\right)$

$$1.545 - 1.645 \times \frac{0.010}{\sqrt{10}} < \mu < 1.545 - 1.645 \cdot \frac{0.010}{\sqrt{10}}$$

$$\Rightarrow \mu \in (1.540; 1.550)$$

* Remark: Nếu n, δ cố định thì độ tin cậy càng cao

→ Khoảng tin cậy (CI) càng dài (longer)

c.) Construct a 95% upper confidence bound on μ:

$$\bar{x} + z_{\alpha} \frac{\delta}{\sqrt{n}}$$

$$= 1.545 + z_{0.05} \times \frac{0.010}{\sqrt{n}}, \quad z_{0.05} = 1.645$$

$$= 1.55$$

d.) Construct a 90% lower confidence bound on μ:

$$\bar{x} - z_\alpha \frac{\sigma}{\sqrt{n}} \; ; \; z_\alpha = z_{0.1}$$

* Remark:

1.) A length of a CI is: $2 \cdot z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}$

2.) The error: $|\bar{x} - \mu|$
(Sai số)

It is clear that: the error $= |\bar{x} - \mu| \leq z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}$

Ex: σ is known, $\sigma = 6$.
(SB8/18)
She desires to be 98% confidence that $|\bar{x} - \mu| \leq 4$

Solve:
$$z_{\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} = 4$$
$$\Leftrightarrow z_{0.01} \times \frac{6}{\sqrt{n}} = 4$$
$$\Rightarrow n = 12.173$$

The required sample size is: $n = 13$.
(Kích thước mẫu cần thiết)

Solution 2:
$$n = \left\lceil \left(z_{\frac{\alpha}{2}} \cdot \frac{\sigma}{E} \right) \right\rceil = \left\lceil \left(\frac{2.326 \times 6}{4} \right)^2 \right\rceil$$
$$= \lceil 12.173 \rceil = 13$$

* Remark: (Interpreting a 95% CI)

If an infinite number of random samples are collected and a 95% CI for μ is computed from each sample, 95% of these intervals will contain μ.

2.) Confidence on μ of a normal distribution, σ is unknown

σ is know	σ is unknow
use μ	use s (sample stadard de..)
use $z_{\frac{\alpha}{2}}$	use $t_{\frac{\alpha}{2}, n-1}$
use z_{α}	use $t_{\alpha, n-1}$

t_{n-1} : t - distribution with $n-1$ degree of freedom

$$t_{0.05,18} = t_{\alpha, n-1} \Rightarrow \begin{cases} \alpha = 0.05 \\ n-1 = 18 \end{cases}$$

a.) A $100(1-\alpha)\%$ two-sided CI on μ is:

$$\bar{x} - t_{\frac{\alpha}{2}, n-1} \cdot \frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2}, n-1} \cdot \frac{S}{\sqrt{n}}$$

b.) A $100(1-\alpha)\%$ upper confidence bound is:

$$\mu \leq \bar{x} + t_{\alpha, n-1} \frac{S}{\sqrt{n}}$$

c.) A $100(1-\alpha)\%$ lower confidence bound is:

$$\mu \geq \bar{x} - t_{\alpha, n-1} \frac{S}{\sqrt{n}}$$

Ex: The population has normal distribution, δ is unknown

A 95% two-sided CI on μ is:

$$\bar{x} - t_{\frac{\alpha}{2}, n-1} \cdot \frac{S}{\sqrt{n}} < \mu < \bar{x} + t_{\frac{\alpha}{2}, n-1} \cdot \frac{S}{\sqrt{n}}$$

$\bar{x} = 3450$, $n = 28$, $S = 600$

Confidence level $= 95\% = 1 - \alpha \Rightarrow \alpha = 0.05$

$\Rightarrow t_{\alpha/2, n-1} = t_{0.025, 27} = 2.052$

$\Rightarrow$ $3218 < \mu < 3682$, or $\mu \in (3218, 3682)$

$\mu \in (3218, 3282)$

3) Confidence interval for p - Khoảng tin cậy cho tỷ lệ p.

p: population proportion (tỷ lệ của tổng thể - chưa biết)

$p = \hat{p} = \frac{x}{n}$ Sample proportion (tỷ lệ mẫu)

If $\begin{cases} np \geq 5 \\ np(1-p) \geq 5 \end{cases}$ then we have

$$Z = \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

Since, $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1 - \alpha$.

$$\Rightarrow P\left(-z_{\alpha/2} \leq \frac{\hat{P} - p}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

$$\Rightarrow P\left(-z_{\alpha/2} \leq \frac{p - \hat{P}}{\sqrt{\frac{p(1-p)}{n}}} \leq z_{\alpha/2}\right) = 1 - \alpha.$$

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

Thứ ngày

a) A $100 \times (1-\alpha)\%$ CI for p is given by:

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

$\hat{p}$: is the proportion of a sample of size n $\left(\hat{p} = \frac{x}{n}\right)$

b) A $100 \times (1-\alpha)\%$ CI upper confidence bound for p is:

$$p \leq \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

c) A $100(1-\alpha)\%$ lower confidence bound for p is:

$$p \geq \hat{p} - z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

VD: Population homeless people

$\hat{p}$: the proportion of homeless people who are veterants

A 95% upper confidence bound for p is:

$$p \leq \hat{p} + z_{\alpha}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$\Rightarrow 250 = n$

$\hat{p} = \frac{x}{n} = \frac{47}{250} = 0.188$

Confidence level $= 1 - \alpha = 95\% \rightarrow \alpha = 0.05$.

⇒ Critical solve: $z_\alpha = z_{0.05} = 1.645$

⇒ $p \leq 0.188 + 1.645 \times \sqrt{\frac{0.188(1-0.188)}{250}}$

⇒ $p \leq 0.029$.

VD: 1000 randomly selected cases of lung cancer, 750 resulted in death within 10 years. Calculate a 95% two-sided confidence interval of the death rate from lung cancer.

$z_\alpha = z_{0.05} = 1.645$

$z_{\alpha/2} = z_{0.025} = 1.96$.

$n = 1000$

$\hat{p} = \frac{x}{n} = \frac{750}{1000} = 0.75$.

A 95% two-sided CI interval of the death rate from lung cancer:

$0.75 - 1.96 \times \sqrt{\frac{0.75 \times (1-0.75)}{1000}} \leq p \leq 0.75 + 1.96 \times \sqrt{\frac{0.75 \times (1-0.75)}{1000}}$

$0.723 \leq p \leq 0.776$.

Problem: Find the required sample size such that the error: $= |\hat{p} - p| \leq E$

Case 1: p is known

Solve: $z_{\alpha/2}\sqrt{\frac{p(1-p)}{n}} = E \Rightarrow n = \left(\frac{z_{\alpha/2}}{E}\right)^2 \times p(1-p)$

$\Rightarrow$ Then choose $n = \left\lceil \left(\frac{z_{\alpha/2}}{E}\right)^2 p(1-p) \right\rceil$

Case 2: p is unknown but $\hat{p}$ is known.

Solve: $z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = E \Rightarrow$ Choose $n = \left\lceil \left(\frac{z_{\alpha/2}}{E}\right)^2 \hat{p}(1-\hat{p}) \right\rceil$

Case 3: p Both p and $\hat{p}$ is unknown

$n = \left\lceil \left(\frac{z_{\alpha/2}}{E}\right)^2 \times 0.25 \right\rceil$

VD: $p = 0.08$ (p is known)

Find min of n you should use to be 99% confident that the point estimate of p (is $\hat{p}$) will be within 0.04 around p.

Solve: $|\hat{p} - p| \leq 0.04$

$\Rightarrow n = \left\lceil \left(\frac{z_{\alpha/2}}{E}\right)^2 p(1-p) \right\rceil = \left\lceil \left(\frac{2.58}{0.04}\right)^2 0.08(1-0.08) \right\rceil$

$= \lceil 306.1944 \rceil = 307$

$E = 0.04$

$z_{\alpha/2} = z_{0.005} = 2.58$

Thứ ngày

VD: (Both p and $\hat{p}$ is unknown) $z_{0.01} = 2.33$

Find n such that $|\hat{p} - p| \leq 6\%$ with $1-\alpha = 98\%$

$$n = \left\lceil \left(\frac{z_{\alpha/2}}{E}\right)^2 \times 0.25 \right\rceil = \left\lceil \left(\frac{\alpha\ 2.33}{0.06}\right)^2 \times 0.25 \right\rceil = \lceil 377.006 \rceil$$

$$= 378$$

VD:

$n = 4000 \quad \rightarrow \hat{p} = \frac{x}{n} = 0.625$

$x = 2500$

$E = 5\% = 0.05$

$1-\alpha = 94\%$

$\rightarrow \alpha = 0.06.$

Chapter 9: Test of hypotheses for a single sample
Kiểm định các giả thuyết cho một mẫu đơn

Statistical hypothesis (Giả thuyết thống kê)

Hypothesis: is a assestion about the parameters θ

θ: μ (mean); p (proportion)

Null hypothesis H_0 (Gthuyết ko) - Gthuyết gốc) $\theta = \theta_0$

Alternative hypothesis H_1: (Gthuyết đối)

10 15 20

HONG HA

Thứ ngày

Case 1: (two-tails test / Two-sided test)

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$$

Case 2: (Left-tailed test / Lower-sided test)

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$$

Case 3: (Right-tailed test / Upper-sided test)

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$$

§ Type of error (các loại sai lầm)

* α = P(type I error) = P(reject H_0 when H_0 is true)

= P($\bar{x}$ is in critical region | $\mu = \mu_0$)

α: the significance level (mức ý nghĩa)

β = P(type II error) = P(Fail to reject H_0 when H_0 is false)

§ Hyprothesis testing procedure for μ (σ is known)

A. Traditional method

1. Case 1 (two-tails test)

$$\begin{cases} H_0 : \mu = \mu_0 \\ H_1 : \mu \neq \mu_0 \end{cases}$$

⊕ Determine a test statistic (thống kê kiểm định) and its value

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

⊕ Determine the critical value (percentage point): $z_{\alpha/2}$

⊕ Make conclusion:

If $z_0 < -z_{\alpha/2}$ or $z_0 > z_{\alpha/2}$ ($|z_0| > z_{\alpha/2}$)

→ Reject H_0

If $-z_{\alpha/2} \leq z_0 \leq z_{\alpha/2}$ $(|z_0| \leq z_{\alpha/2})$

→ Fail to reject H_0
(Do not reject H_0)

2.) Case 2: (Left-tailed test, lower-tailed test)

$$\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$$

+) The critical value: z_α

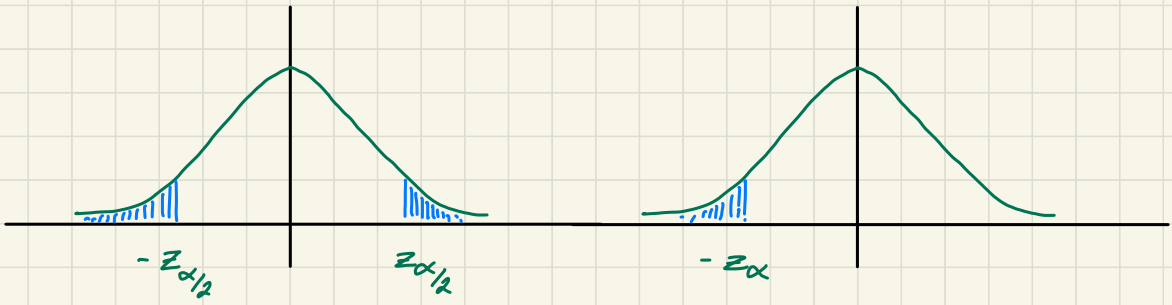
+) Conclusion: If $z_0 < -z_\alpha$, the reject H_0

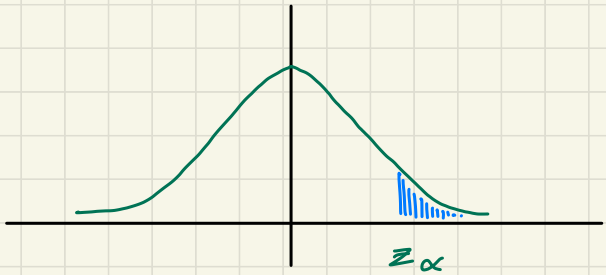
If $z_0 \geq -z_\alpha$ → Fail to reject H_0

3.) Case 3: $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 \end{cases}$

If $z_0 > z_\alpha$ → Reject to H_0

If $z_0 \leq z_2$ → Fail to reject H_0





Ex1: Population is normally distributed
(Sl 8/18)

$\sigma = 0.03$ (σ is known)
$n = 43$, $\bar{x} = 1.64$, the significiance level is $\alpha = 0.05$

⊕ The hypotheses: $\begin{cases} H_0: \mu = 1.7 \\ H_1: \mu \neq 1.7 \end{cases}$ $(\mu_0 = 1.7)$

⊕ The test statistic is

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{1.64 - 1.7}{0.03/\sqrt{43}} = -13.115$$

(Note: In this problem, we will reject H_0 if $z_0 < -z_{\alpha/2}$ or $z_0 > z_{\alpha/2}$)

⊕ The critical value: $z_{\alpha/2} = z_{0.025} = 1.96$

⊕ Since $z_0 = -13.115 < -z_{\alpha/2} = -1.96$
$\Rightarrow$ Reject H_0 and accept H_1

✳ $\begin{cases} H_0: \mu = \dots \\ H_1: \mu \begin{matrix} \neq \\ > \\ < \end{matrix} \end{cases}$ (not $\geq$ or $\leq$)

Ex 2: Population is normally distributed
(Sl 12/18)

$\sigma = 0.03$ (σ is known)

$n = 43$, $\bar{x} = 1.64$, the significiance level is $\alpha = 0.05$

⇒ The hypotheses: $\begin{cases} H_0: \mu = 1.6 \\ H_1: \mu > 1.6 \end{cases}$ ($\mu_0 = 1.6$)

⇒ The test statistic:

$$z_0 = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{1.64 - 1.6}{0.03/\sqrt{43}} = 8.743$$

⇒ The critical value: $z_\alpha = z_{0.05} = 1.645$

⇒ $z_0 = 8.743 > z_\alpha = 1.645 \rightarrow$ Reject H_0 (Accept H_1)

⇒ We conclude that the mean height is greater than 1.6 m!

B.) P_ value method

1.) Two-tailed test: $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu \neq \mu_0 \end{cases}$

$$P\text{_value} = P(Z < -|z_0| \text{ or } Z > |z_0|)$$

$$= 2 \cdot P(Z < -|z_0|) = 2 \cdot P(Z > |z_0|)$$

Conclusion: ⇒ If P_ value $\leq \alpha$, then reject H_0

⇒ If P_ value $> \alpha$, then fail to reject H_0

2.) Left-tailed test: $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu < \mu_0 \end{cases}$

$$P\text{-value} = P(Z < z_0)$$

3.) Right-tailed test: $\begin{cases} H_0: \mu = \mu_0 \\ H_1: \mu > \mu_0 \end{cases} \Rightarrow P\text{-value} = P(Z > z_0)$

Ex (SL 12/18)

$z_0 = 8.743$

$\begin{cases} H_0: \mu = 1.6 \\ H_1: \mu > 1.6 \end{cases}$

$$P\text{-value} = P(Z > z_0) = P(Z > 8.743)$$
$$= 1 - P(Z < 8.743) \approx 1 - 1 = 0$$

Thus $P\text{-value} = 0 < \alpha = 0.05 \rightarrow$ Reject H_0

§ Test of hypotheses for population mean μ (σ is unknown)

σ is known	σ is unknown
The test statistic: $z_0 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$	The test statistic: $t_0 = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$
use σ	use s
use $z_{\alpha/2}$	use $t_{\alpha/2, n-1}$
use z_α	use $t_{\alpha, n-1}$

Note: $n-1$ = degrees of freedom

Ex: (σ is unknown)
Population is normally distributed

$n = 49$, $\bar{x} = 5.1$, $s = 1.2$

The significance level is $\alpha = 0.05$

Đ The value of the test statistic is:

$$t_0 = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{5.1 - 4.5}{1.2/\sqrt{49}} = 3.5$$

Đ $\begin{cases} H_0 : \mu = 4.5 \\ H_1 : \mu > 4.5 \end{cases}$

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

Note: Since $t_0 = 3.5 > t_{\alpha, n-1} = 1.677 \longrightarrow$ Reject H_0, accept H_1

§ Test of hypotheses for population proportion p

* Tương tự với bài toán kiểm định cho μ khi σ đã biết, lưu ý là:

The test statistic: $Z_0 = \dfrac{\hat{P} - P_0}{\sqrt{\dfrac{P_0(1-P_0)}{n}}}$

Ex (Slide 17)

Test the hypotheses: $\begin{cases} H_0 : P = 3\% = 0.03 \\ H_1 : P < 0.03 \end{cases}$ $(P_0 = 0.03)$

Đ The test statistic is:

$$Z_0 = \frac{\hat{P} - P_0}{\sqrt{\dfrac{P_0(1-P_0)}{n}}} = \frac{\dfrac{6}{135} - 0.03}{\sqrt{\dfrac{0.03(1-0.03)}{135}}} = 0.984$$

$z_\alpha = z_{0.05} = 1.645$

Because $z_0 > z_\alpha \rightarrow$ Fail to reject H_0

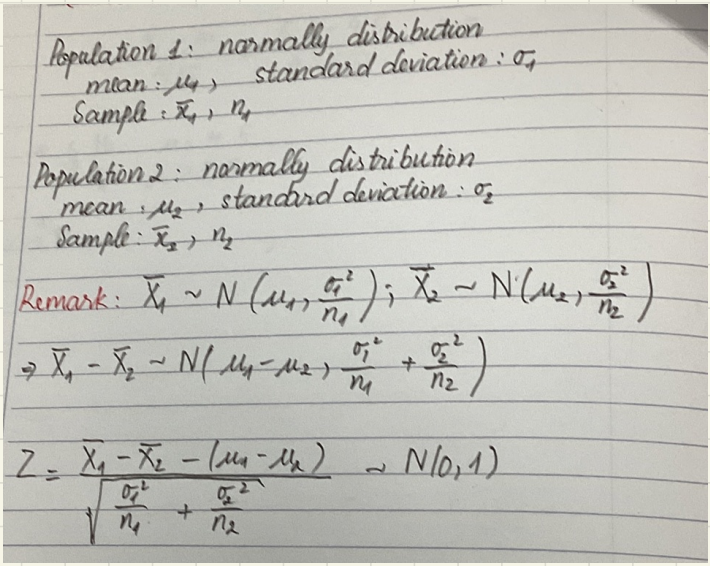
Solution (Use P_value)

P_value $= P(z < z_0) = P(z < 0.984) = 2.144$

Since P_value $> \alpha \rightarrow$ Fail to reject H_0

Chapter 10:

Statical interence for Two Samples



I. Interence on the difference in means of two normal distribution variances known. $(\mu_1 - \mu_2)$ ↗

1. CI on $\mu_1 - \mu_2$

(a) A $100(1-\alpha)\%$ two-sided CI on $\mu_1 - \mu_2$ is given by

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

(b) A $100(1-\alpha)\%$ upper confidence bound on $\mu_1 - \mu_2$ is:

$$\mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

(c) A $100(1-\alpha)\%$ lower confidence bound on $\mu_1 - \mu_2$ is

$$\mu_1 - \mu_2 \geq (\bar{x}_1 - \bar{x}_2) - z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma^2}{n_2}}$$

Ex 1: Population 1: $\sigma_1^2 = 1.5$ Population 2: $\sigma_2^2 = 1.2$

$n_1 = 13$ $n_2 = 20$

$\bar{x}_1 = 89.6$ $\bar{x}_2 = 92.5$

Confidence level: $1 - \alpha = 95\% \Rightarrow \alpha = 5\% = 0.05$

A 95% Confidence interval on the difference in mean $(\mu_1 - \mu_2)$ is:

$$(\bar{x}_1 - \bar{x}_2) - z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \mu_1 - \mu_2 < (\bar{x}_1 - \bar{x}_2) + z_{\alpha/2} \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$\Rightarrow -3.684 < \mu_1 - \mu_2 < -2.116$$

Or $\mu_1 - \mu_2 \in (-3.684, -2.116)$

b) A 99% lower confidence bound on $\mu_1 - \mu_2$ is

$$\mu_1 - \mu_2 \geq (\bar{x}_1 - \bar{x}_2) - z_\alpha \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}$$

$z_\alpha = z_{0.01} = 2.326$

2. Test of hypotheses for difference in means $\mu_1 - \mu_2$.

The test statistic is $z_0 = \dfrac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_1^2}{n_2}}}$

Case 1: $\begin{cases} H_0: \mu_1 - \mu_2 = \Delta_0 \\ H_1: \mu_1 - \mu_2 \neq \Delta_0 \end{cases}$ (two-tailed test)

If $z_0 < -z_{\alpha/2}$ or $z_0 > z_{\alpha/2}$ → Reject H_0

⇒ P_value $= 2P(Z < -|z_0|) = 2\underline{P}(Z > |z_0|)$

Case 2: $\begin{cases} H_0: \mu_1 - \mu_2 = \Delta_0 \\ H_1: \mu_1 - \mu_2 < \Delta_0 \end{cases}$ (Left-tailed test / lower ——)

⇒ Reject H_0 if $z_0 < -z_\alpha$

⇒ P_value $= \underline{P}(Z < z_0)$

Case 3: $\begin{cases} H_0: \mu_1 - \mu_2 = \Delta_0 \\ H_1: \mu_1 - \mu_2 > \Delta_0 \end{cases}$ (Right-tailed test)

⇒ Reject H_0 if $z_0 > z_\alpha$

⇒ P_value $= \underline{P}(Z > z_0)$

Ex: Formulation 1: $\sigma_1 = 8$, $n_1 = 10$, $\bar{x}_1 = 121$. (μ_1 is unknow)

Formulation 2: (has a new drying ingredient) (μ_2 is unknown)

$\sigma_2 = 8$, $n_2 = 10$, $\bar{x}_2 = 112$

Siguificant level : $\alpha = 0.05$

The hypotheses: $\begin{cases} H_0 : \mu_1 - \mu_2 = 0 \\ H_1 : \mu_1 - \mu_2 > 0 \end{cases}$ $(\Delta_0 = 0)$

The value of the test statistic is:

$$z_0 = \frac{(\bar{x}_1 - \bar{x}_2) - \Delta_0}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_1^2}{n_2}}} = 2.516$$

⊕ $z_\alpha = z_{0.05} = 1.645$

⊕ Since $z_0 > z_\alpha$ → Reject H_0, accept H_1.

→ We conclude that adding the new gredient reduce the drying time.

Remark: P_ value $= P(Z > z_0) = P(Z > 2.516)$

$= 1 - P(Z < 2.516) = 0.006$

Since p_ value $= 0.006 < \alpha = 0.05$

→ Reject H_0

II. Inference on $\mu_1 - \mu_2$ of two normal distributions

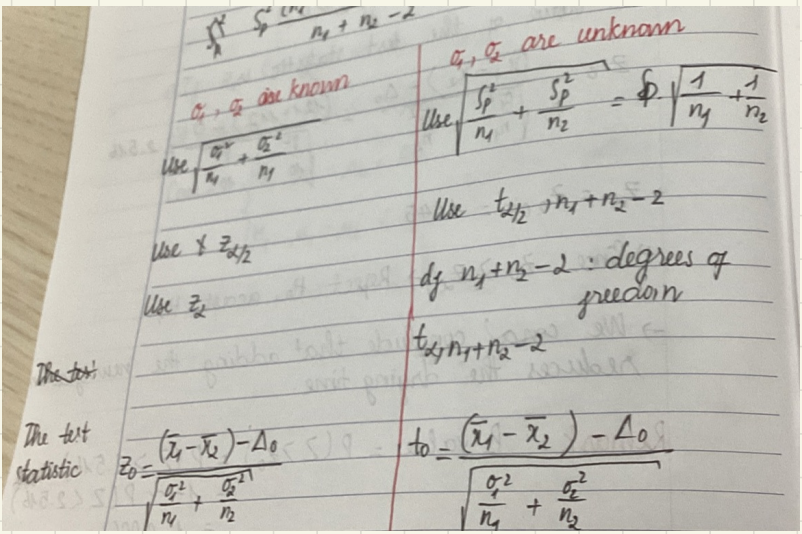
(Assume σ_1, σ_2 are unknown and $\sigma_1 = \sigma_2$)

Sample 1: n_1, $\bar{x}_1$, s_1 (s_1 is sample standard deviation)

Sample 2: n_2, $\bar{x}_2$, s_2

Def: Pooled variance (phương sai gộp)

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



Ex 1: Two population are normally distributed and have the same variance

Catalyst 1: $n_1 = 8$, $\bar{x}_1 = 92.255$, $s_1 = 2.39$

Catalyst 2: $n_2 = 8$, $\bar{x}_2 = 92.733$, $s_2 = 2.98$

Ⓓ Pooled variance:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 7.3$$

A 95% Confidence interval for difference in means $\mu_1 - \mu_2$

$$(\bar{x}_1 - \bar{x}_2) - t_{\alpha/2, n_1+n_2-2}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}} < \mu_1 - \mu_2 <$$

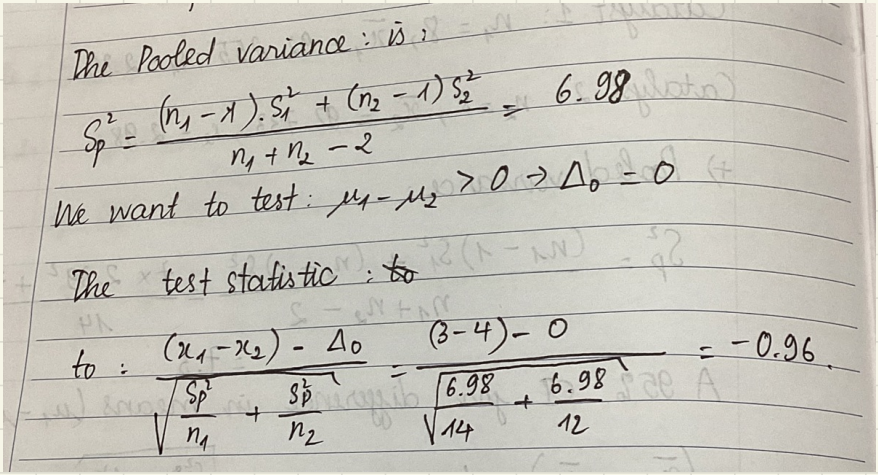
$$(\bar{x}_1 - \bar{x}_2) + t_{\alpha/2,\, n_1+n_2-2}\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}$$

We have $t_{\alpha/2,\, n_1+n_2-2} = t_{0.025,\, 14} = 2.145$

$$\Rightarrow -3.376 < \mu_1 - \mu_2 < 2.42$$

Ex2: Sample 1: $n_1 = 14$, $\bar{x}_1 = 3$, $s_1 = 2.5$

" 2: $n_2 = 12$, $\bar{x}_2 = 4$, $s_2 = 2.8$



The Pooled variance is:

$$s_p^2 = \frac{(n_1 - 1).s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 6.98$$

We want to test: $\mu_1 - \mu_2 > 0 \Rightarrow \Delta_0 = 0$

The test statistic: ~~to~~

$$t_0 : \frac{(x_1 - x_2) - \Delta_0}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_2}}} = \frac{(3-4) - 0}{\sqrt{\frac{6.98}{14} + \frac{6.98}{12}}} = -0.96.$$