January 7, 2023

DSML : CC Fundamentals

Hypothesis Testing - 3

$M_w \approx 80$

$M_e \approx 53.$

→ Q] Do fewer people attend class on weekends?

Recap:

* Null Hypothesis : H_0
* Alternate Hypothesis : H_a
* p-value.
* Test statistic
* Significance level: α.
* z-test
* t-test (1-sample).

Class begins @ 9:05 p.m.

TERMINOLOGY
EVERYWHERE!

Hypothesis testing Framework.

1] Set up the Null and Alternate hypothesis.

2] Choose the correct test statistic —

3] Select a type of test: left-tailed, right-tailed or two-tailed.

4] Compute the p-value —

5] If p-value is less than α, the significance level, then reject the null hypothesis.

Recap

$\mu = 1800$
$\sigma = 100$

A retailer has 2000 stores in the country

Historical data tells us that weekly sales of shampoo bottles has an average of 1800, with a standard deviation of 100

Sales team wants to improve sales by hiring a marketing team

Hiring a marketing team can be expensive, so we need to be very sure that they will improve sales

Before deploying their strategy for all 2000 stores, they are tested in 50 stores

On the 50 stores, their average sales for that week was 1850

You are the data scientist who should tell your sales team whether this is statistically significant

Sales team has said that we will hire only if we are 99 % confident $\alpha = 0.01$

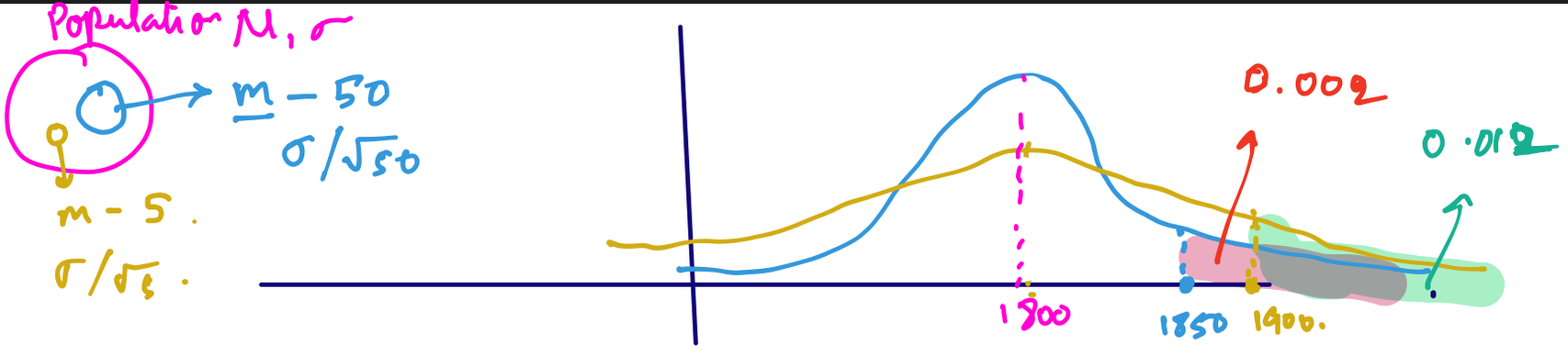
Another marketing team is also being considered

They are tested on 5 stores

On the 5 stores, their average sales for that week was 1900

Would you say this team is better than the first one?

Between the “blue team” and the “yellow team”, whom will you choose?



Recap

$\alpha = 0.01$

$\mu = 1800$

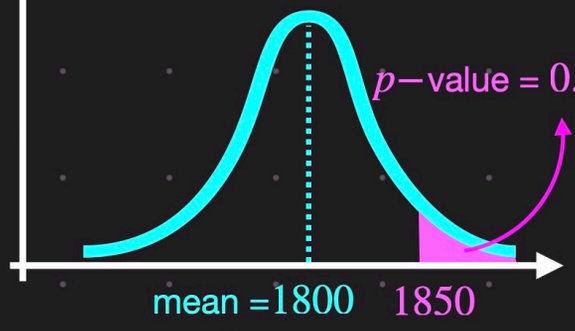
$\sigma = 100$

50 stores with average of 1850

$H_0 : \mu_b = 1800$

$H_a : \mu_b > 1800$

Distribution of m_b under the assumption that H_0 is true



std dev = $\frac{100}{\sqrt{50}}$

$z = \frac{1850 - 1800}{100/\sqrt{50}} = 3.53$

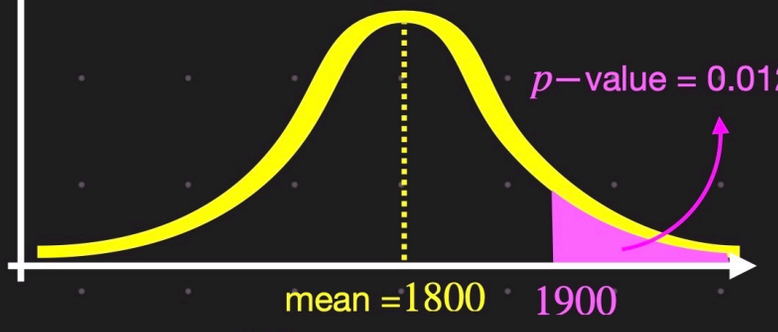
Reject H_0

5 stores with average of 1900

$H_0 : \mu_y = 1800$

$H_a : \mu_y > 1800$

Distribution of m_y under the assumption that H_0 is true



std dev = $\frac{100}{\sqrt{5}}$

$z = \frac{1900 - 1800}{100/\sqrt{5}} = 2.23$

Fail to reject H_0

Recap: Sachin Tendulkar's batting.

→ Average ≈ 43 → σ ≈ 42

Q1] Does Sachin bat better in the first innings vs. the second?

→ first ≈ 46
second ≈ 40
→ Not significant.

Q2] Does Sachin score more when India wins vs. loses?

win ≈ 51
loss ≈ 35.
→ significant.

Recap : t - test vs. z - test.

$$z = \frac{m - \mu}{\sigma/\sqrt{n}} \qquad t = \frac{m - \mu}{s/\sqrt{n}}$$

μ → population mean.

→ σ → Population std

→ s → Sample std.

Drug Recovery

Suppose 2 companies develop drugs to cure a disease.

Drug 1 was tested on 100 people, and the recovery days look like:

→ [8, 5, 9, 10, ..., 16] , μ_1 = 7.1 days.

Drug 2 was tested on 120 people, and the recovery days look like:

— [12, 4, 7, 13, ..., 8] , μ_2 = 8.07 days.

H_0: $\mu_1 = \mu_2$

H_a: $\mu_1 < \mu_2$

Can we say one drug is better/worse than the other?

Drug Recovery : Hypothesis testing.

$\mu_b \rightarrow$ 1st drug
$\mu_y \rightarrow$ 2nd drug.

1] $H_0 : \mu_b = \mu_y$
$H_a : \mu_b > \mu_y$ " Right-tailed 2-sample test."

2] $H_0 : \mu_b = \mu_y$
$H_a : \mu_b < \mu_y$ " Left-tailed 2-sample test"

3] $H_0 : \mu_b = \mu_y$
$H_a : \mu_b \neq \mu_y$ " 2-tailed 2-sample test"

→ 2-sampled t-test.

test statistic

$$\frac{\mu_1 - \mu_2}{\sqrt{s^2\left(\frac{1}{n_1^2} + \frac{1}{n_2^2}\right)}}$$

s → sample std.

IQ across two schools.

Suppose there are two schools competing against each other. Each school says that their students have a higher IQ. So, we conduct a test.

Say the first school had numbers like:

→ [101, 115, 90, ..., 112, 97], $\mu_1 = 103.7$:

Say the second school had numbers like:

→ [108, 105, 109, ..., 111, 93], $\mu_2 = 102.9$.

H_0: $\mu_1 = \mu_2$

H_a

✓ (1) $\mu_1 > \mu_2$ → right

(2) $\mu_1 < \mu_2$ - left

✓ (3) $\mu_1 \neq \mu_2$. → 2

$103.7 > 102.9$

$\mu_1 > \mu_2$

1 or 3 are correct.

Summary: 1-sample test vs. 2-sample test.

	1-Sample	2-Sample
Supply Chain	Compare avg. sales to a fixed value. $H_0: \mu_b = 1800$ $H_a: \mu_b > 1800$	Directly compare the two strategies. $H_0: \mu_1 = \mu_2$ $H_a: \mu_1 \neq \mu_2$
Drug Recovery	Compare recovery time to pop. mean. $\mu_p = 8$ days. observed $\rightarrow m$ $H_0: m = 8$. $H_a: m < 8$.	Directly compare the two drugs. μ_1 μ_2. $H_0: \mu_1 = \mu_2$ $H_a: \mu_1 \neq \mu_2$.
IQ test.	Compare IQ test results with pop. mean. $\mu_p = 100$ observed $\rightarrow m$ $H_0: m = 100$ $H_a: m < 100$, $m > 100$	Directly compare IQ test results of schools. μ_1 μ_2 $H_0: \mu_1 = \mu_2$ $H_a: \mu_1 \neq \mu_2$, $\mu_1 > \mu_1$.

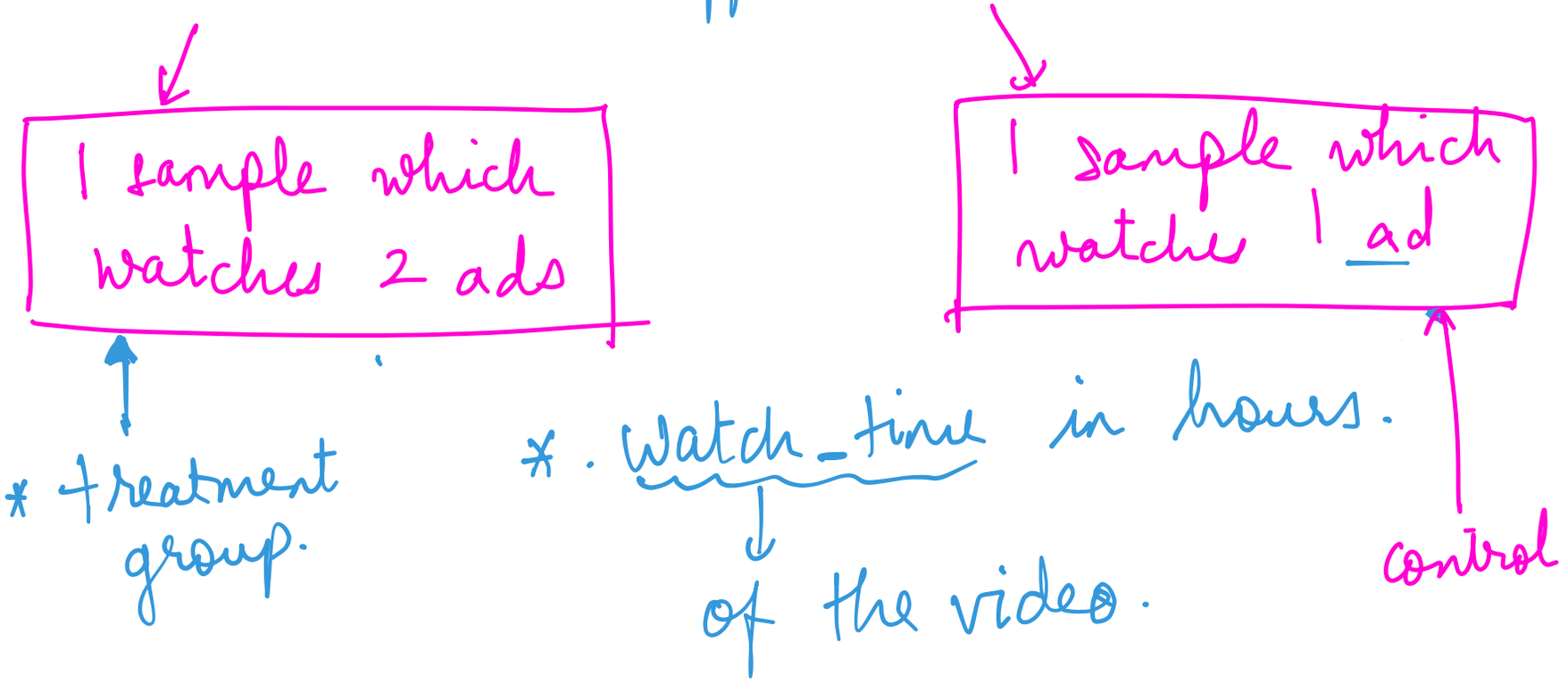
Sachin Tendulkar's Batting : KS - test .

Youtube ads:

Youtube wants to increase its ad revenue. They decide to put 2 ads instead of 1.

Is this a good move? → For them, yes & for us, no!

can we test this effect?



1 sample test

↓

(1) Do this when you have
- population mean,
- population std.

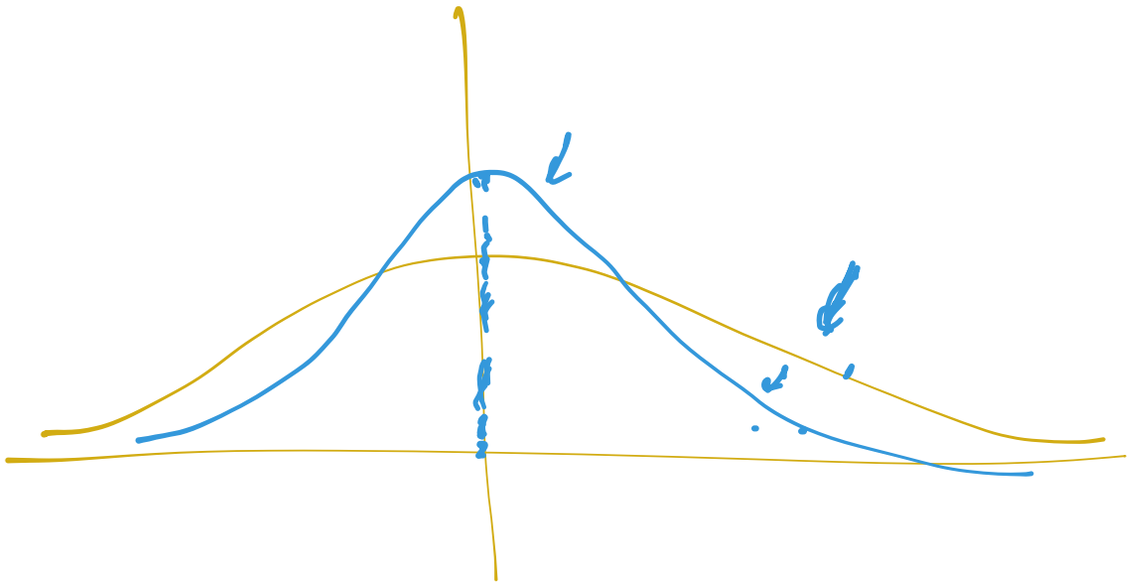
(2) Compare sample mean to population mean.

2-sample test.

↓

(1) Do this when you have 2 samples and no data about pop. μ, σ.

(2) We compare 2 sample means.



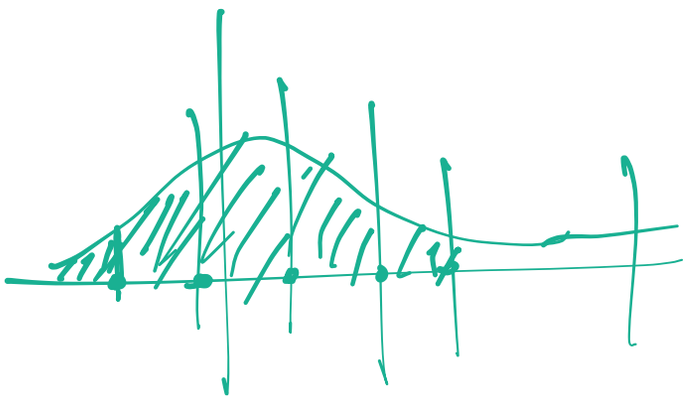
$H_0 : \mu_1 = \mu_2$.

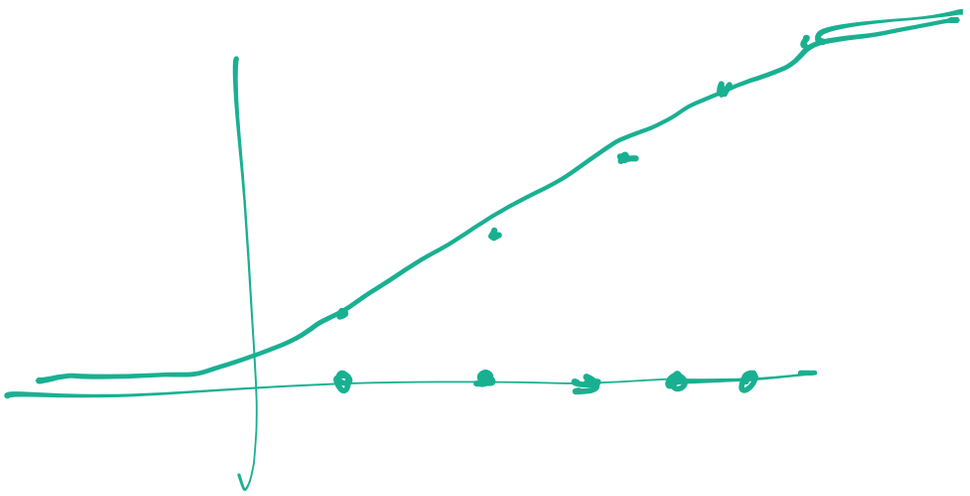
→ 2-sample test.

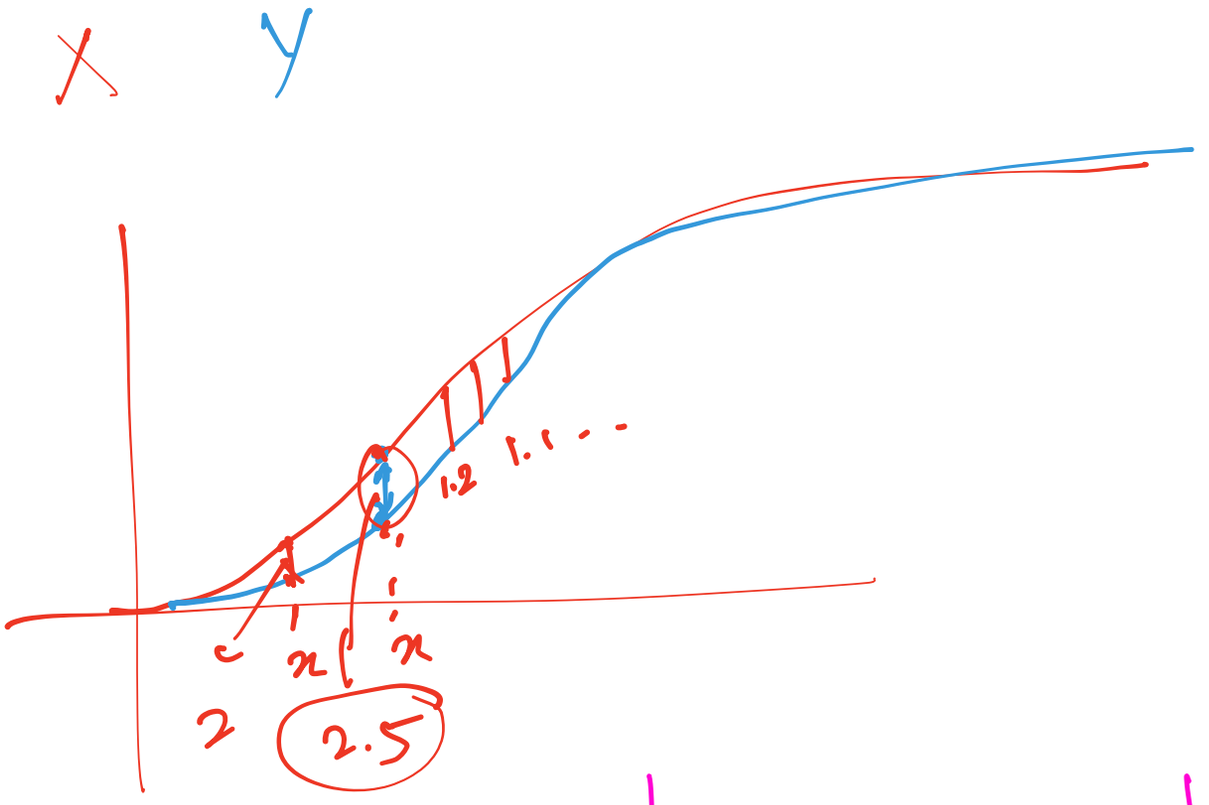
K-S test → test which helps us differentiate between distributions.

K-S test : To check whether 2 Random variables have different distributions or not.

① Calculate the Empirical CDF.







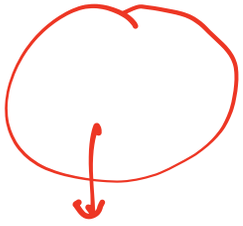
K-S test : $\max \left| cdf(Y) - cdf(X) \right|$

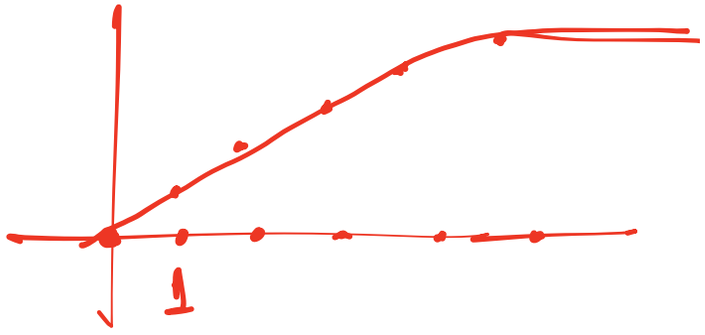
K.S. Test: X, Y → R.V.s

Do they have the same distribution?

H_0 : Yes, X & Y have the same dist.

H_a : No, X & Y do not have the same dist.





$a = [1, 2, 3, 4, 5]$.

$\text{sum}(a) = 15$

$\frac{1}{15}, \frac{2}{15}, \frac{3}{15}, \frac{4}{15}, \frac{5}{15}$.

Recap: Two concepts!

I] 1 sample test vs. 2 sample test.

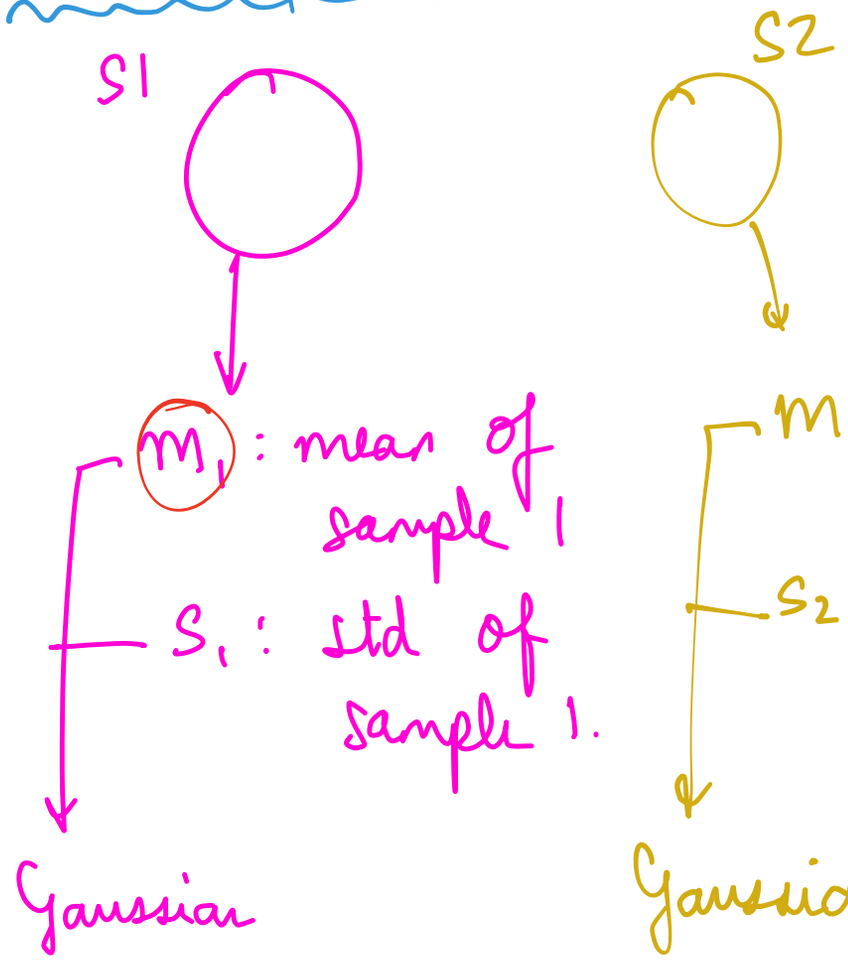
1 sample test:
(a) Compares the sample mean to pop. mean
(b) Done when we have pop. μ

2 sample test:
(a) Compares 2 sample means.
(b) Done when we only have samples.

II] K-S test: (a) Tests whether the ECDFs of two samples is similar or not.

(b) Null: The distributions are the same.

2-sample test.



$$Y = m_1 - m_2$$

$$E[Y] = E[m_1] - E[m_2] = \mu_1 - \mu_2$$

(μ_1: population mean of S1; μ_2: pop. mean of S2.)

$$\text{std}(Y) = \sigma_1/\sqrt{n_1} + \sigma_2/\sqrt{n_2}$$

$$\text{Var}(x) = E[x^2] - (E[x])^2 = \sigma^2.$$

$$\text{Var}(m_1) = E[m_1^2] - (E[m_1])^2 = \sigma_1^2$$

$$\text{Var}(m_1) = \frac{\sigma_1^2}{n}. \qquad \text{Var}(m_2) = \frac{\sigma_2^2}{n}$$

$$\text{std}(m_1) = \frac{\sigma_1}{\sqrt{n_1}} \qquad \text{std}(m_2) = \frac{\sigma_2}{\sqrt{n_2}}$$

$$Y = m_1 - m_2$$

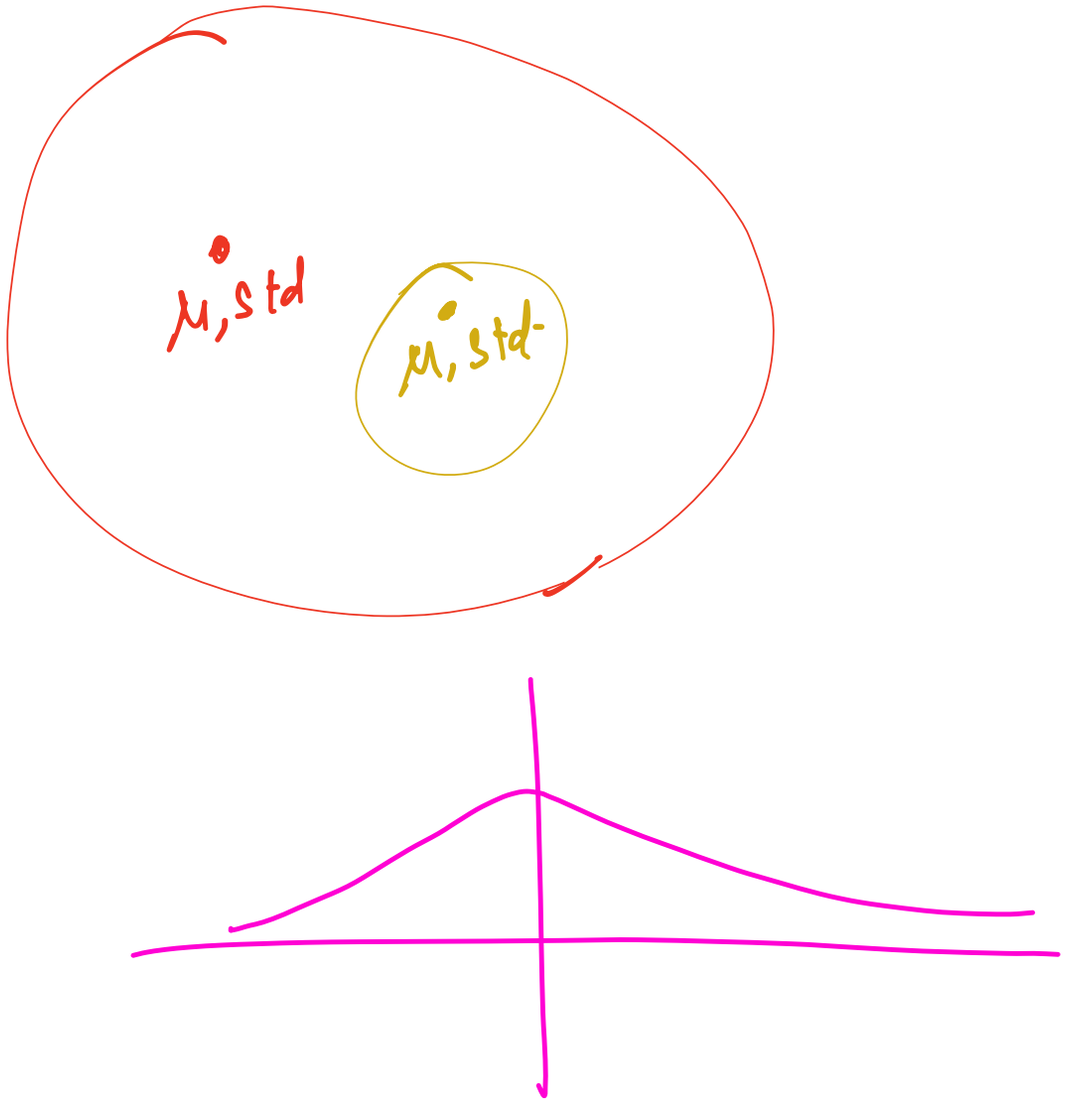
$$\begin{aligned} \text{Var}(Y) &= \text{Var}(m_1 - m_2) \\ &= \text{Var}(m_1) + \text{Var}(m_2) \\ &= \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}. \end{aligned}$$

$$\text{std}(\bar{Y}) = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

Assume: $\sigma_1^2 = \sigma_2^2 = \sigma^2$.

$$\rightarrow \sqrt{s^2\left(\frac{1}{n_1^2} + \frac{1}{n_2^2}\right)}$$

test-statistic: $\dfrac{m_1 - m_2}{\sqrt{s^2\left(\frac{1}{n_1^2} + \frac{1}{n_2^2}\right)}}$


μ, std
μ, std