January 21, 2023

DSML: CC Fundamentals.

Experiment Design

Recap:

(a) Hypothesis testing - α, p-value, Null, Alternate.
(b) z-test } 1 sample, 2 sample, independent, paired,
(c) t-test } left-tailed, right-tailed, 2-tailed.
(d) K-S test
(e) ANOVA
(f) Kruskal
(g) Pearson r
(h) χ^2 test

Class begins @ 9:05 p.m.

Agenda:

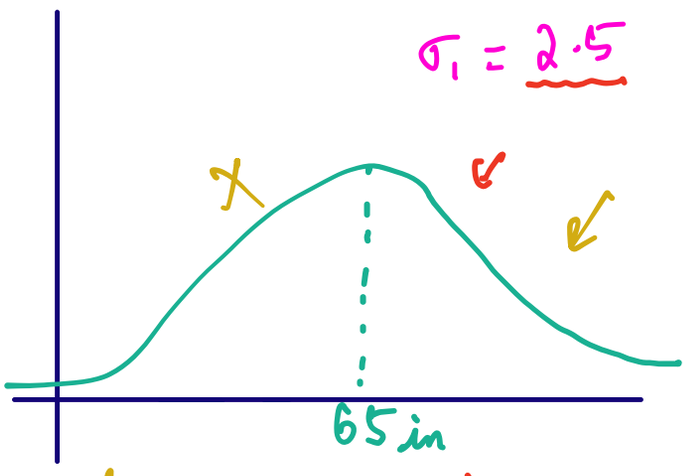
Part 1: Normalization and Standardization.

* sklearn.preprocessing - StandardScaler, MinMaxScaler.

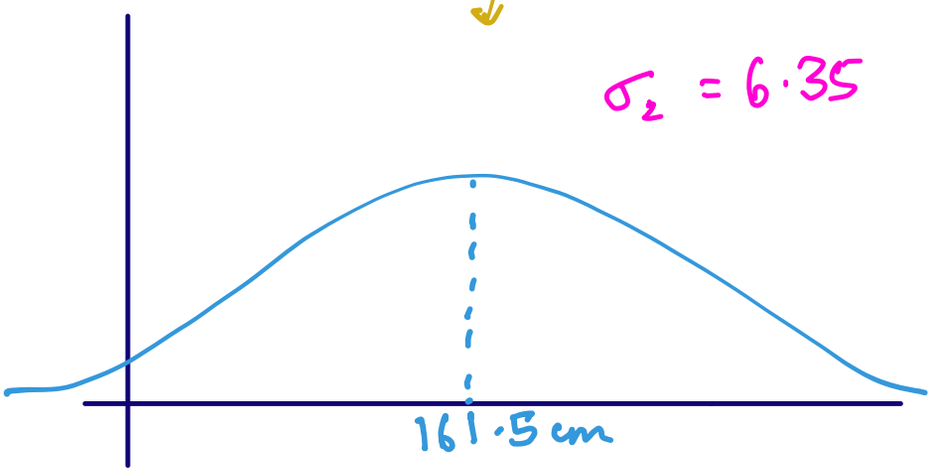
Part 2: Experiment Design.

* Defining α - significance level.
* Defining β - power of the test.
* Obtaining sample size.

Standardization : (Column Standardization)



Height (in)



Height (cm)

$1 inch = 2.54 cm.$

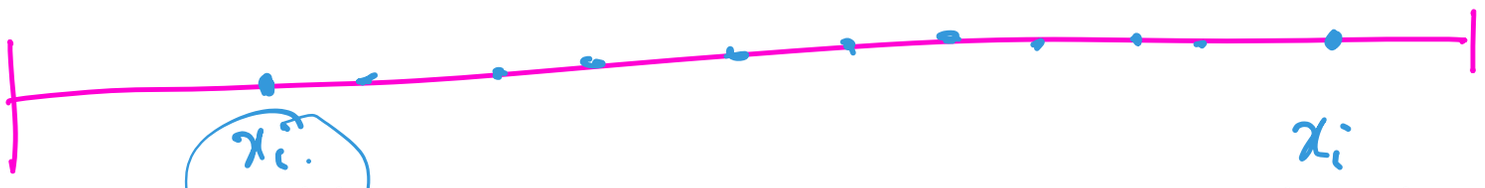
$$X = \{ x_i \in \mathbb{R} \}_{i=1}^{n} \longrightarrow Y = \{ y_i \in \mathbb{R} \}_{i=1}^{n}$$

$$y_i = \frac{x_i - \mu}{\sigma} \rightarrow \text{Standardization.}$$

Normalization.

$$X = \{ x_i \in \mathbb{R} \}_{i=1}^{n} \longrightarrow Y = \{ y_i \in \mathbb{R} \}_{i=1}^{n}$$

$$y_i = \frac{x_i - \min(x_i)}{\max(x_i) - \min(x_i)} = 1$$

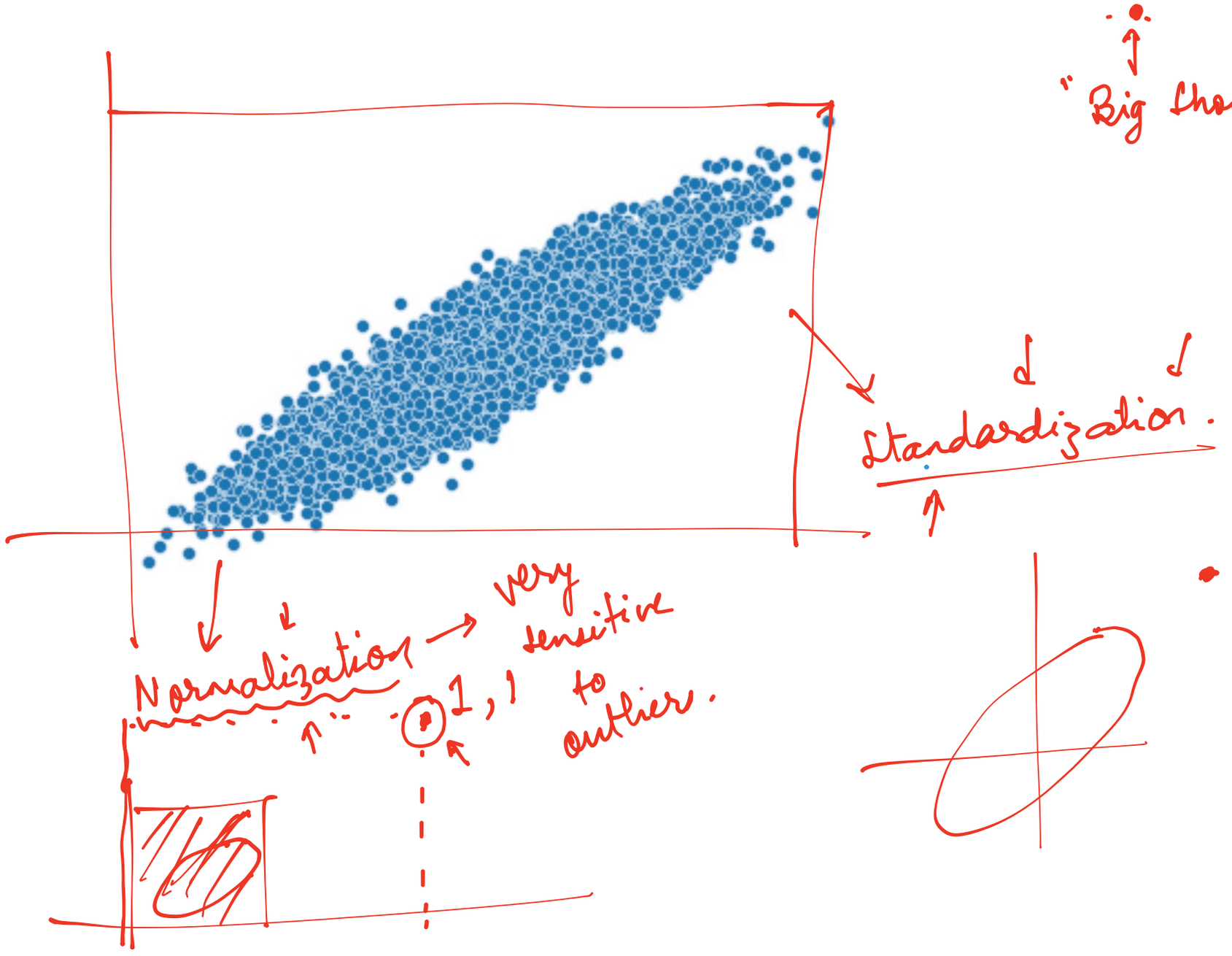


min after normalization

0

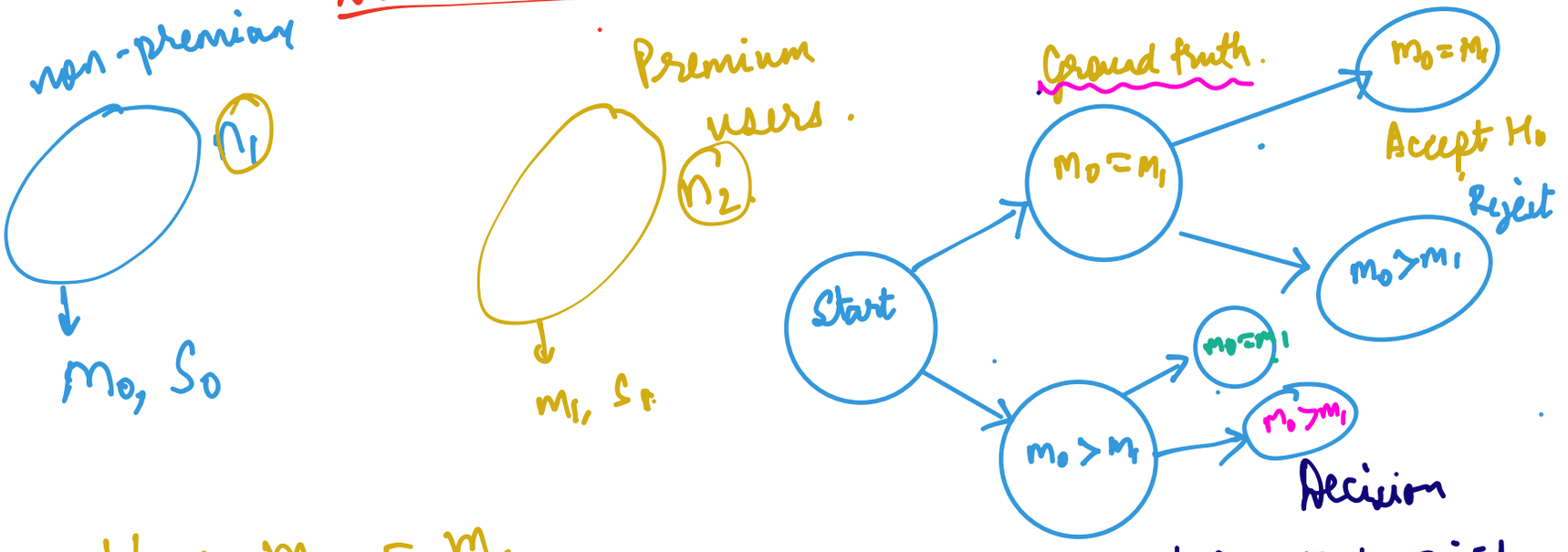
max after normalization

1


"Big Show"
Standardization.
Normalization → very sensitive to outlier.
1, 1

Experiment Design: Youtube data.

Watch time → mins. Premium → 1 or 0



$H_0 : m_0 = m_1$

$H_a : m_0 > m_1$

Reality	Accept	Reject
True	True -ve. $1-\alpha$	False +ve. α
False	False -ve. β	True +ve. $1-\beta$

- α → Significance level. → Type I error.
- β → Measure of False negatives → Type 2 error.

$1-\beta$ → "Power of the test"

Reality	Accept	Reject
True	True -ve. $1-\alpha$	False +ve. α
False	False -ve. β	True +ve. $1-\beta$

$H_0 : m_0 = m_1$ → Right-tailed test.

$H_a : m_0 > m_1$

① $\alpha = 0.01$ ✓ more strict.

vs

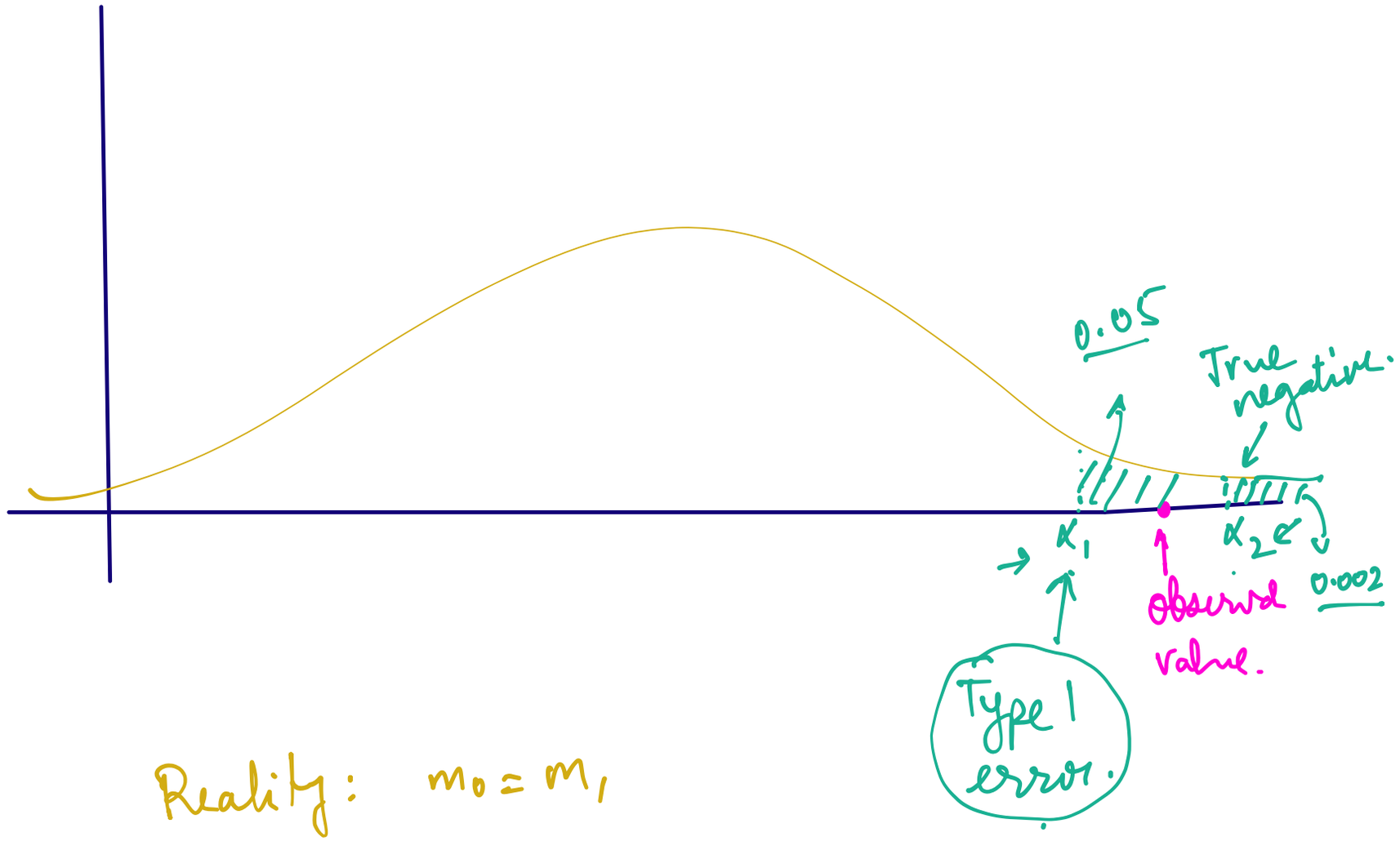
② $\alpha = 0.05$.

Assume Reality is H_0.

99.8

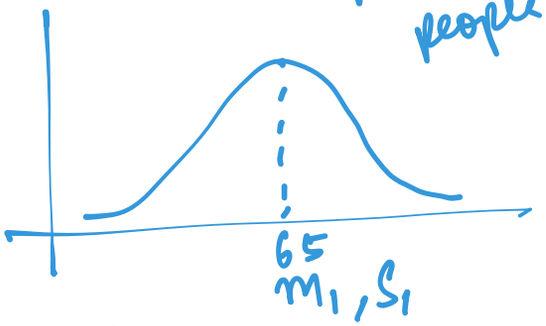
0.002

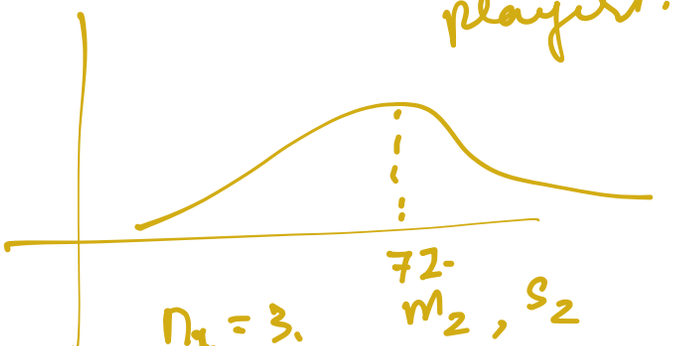
False +ve ; Type I error.



Reality: $m_0 = m_1$

Heights.





$n_1 = 2$

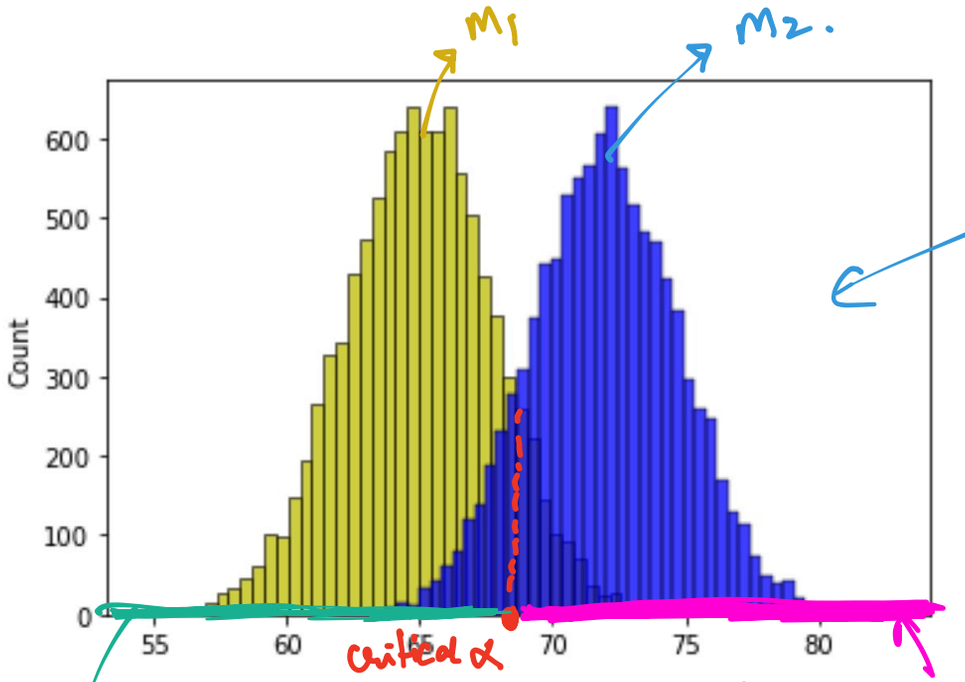
$n_1 = 200$ t-statistic : $\dfrac{m_1 - m_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$ $\leftarrow$ $n_2 = 3$, $n_2 = 300$ $\rightarrow$ $\leftarrow m$.

s = pooled sample standard deviation.

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

$t_1 \rightarrow$ where $n_1 = 2$, $n_2 = 3$ $=$ 2.0.

$t_2 \rightarrow$ —



m_1

m_2.

Reality : $m_1 < m_2$.

$n = 50$ samples.

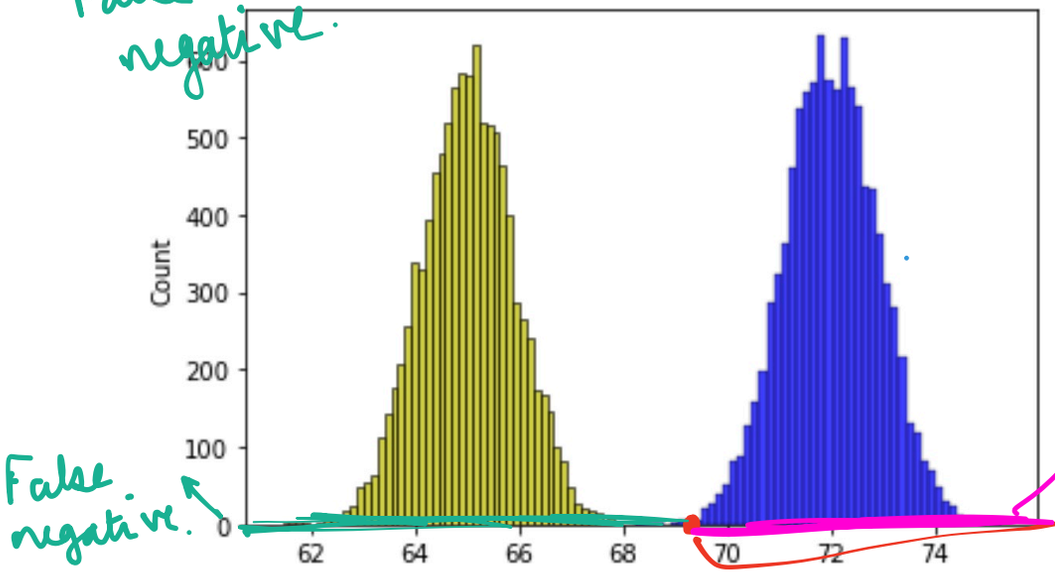
• → observed m_2.

$\sigma / \sqrt{n}$

$n = 50$ samples.

True positive.

False negative.



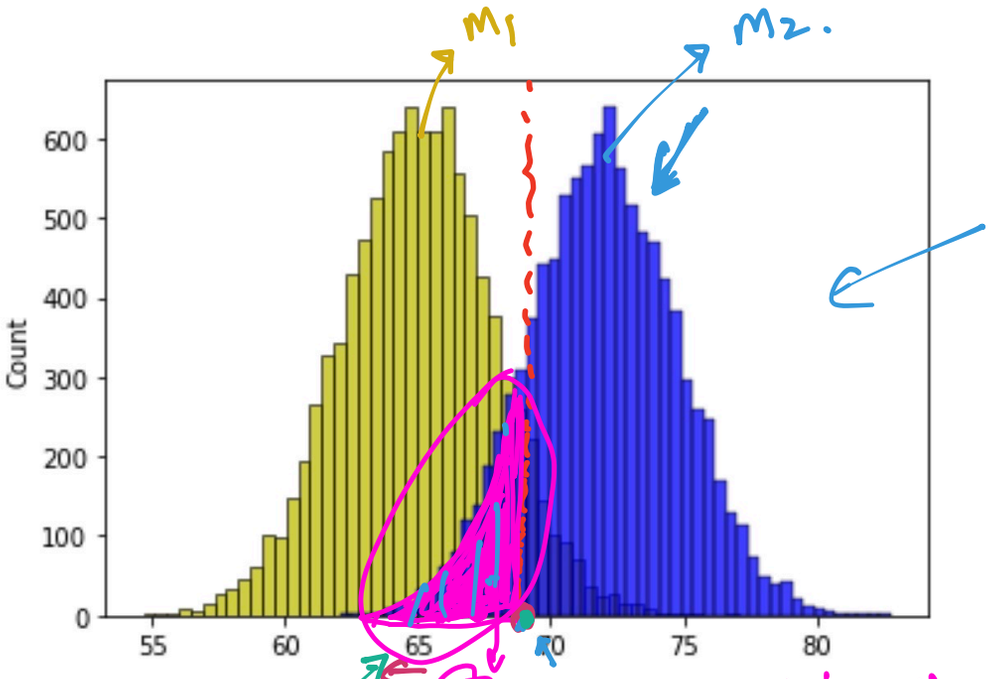
False negative.

True positive.

α.

n ↑ ↓ es. $n = 500$

Reduces False negatives.



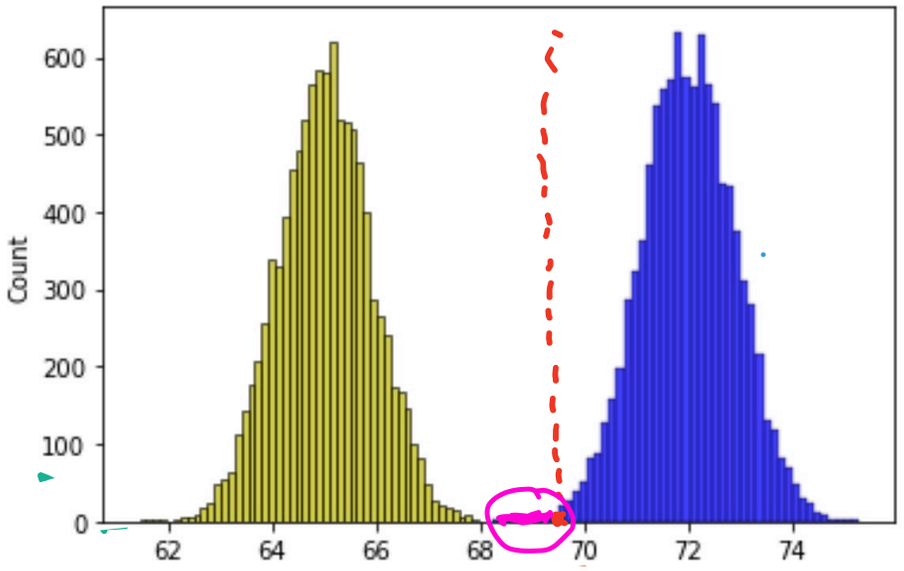
Reality : $m_1 < m_2$.

$n = 50$ samples.

① → observed m_2.

$\sigma / \sqrt{n}$

β → Probability that $m_1 < m_2$ but we fail to reject.



n ↑ ⇒ β ↓. $n = 500$

Reduces False negatives.

False - ve.

$n \rightarrow$ Sample size is the most critical parameter.

Choosing n as large as possible is the best case scenario!

* Core question: What n is just enough??

$$t\text{-Stat} = \frac{m_1 - m_2}{s\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$n_1 = n_2$

$$t_{critical} = \frac{m_1 - m_2}{s\sqrt{2/n}}$$

$m_1 - m_2$ → Error or difference stat.

E → 0.1 s.

decided by α & $1-\beta$.

$t_{critical}$ ~~$t_{critical}$~~ $= \frac{E}{s \cdot \sqrt{2/n}}$

t_c

Cohens' effect size.

$$t_c \times s \times \sqrt{\frac{2}{n}} = E.$$

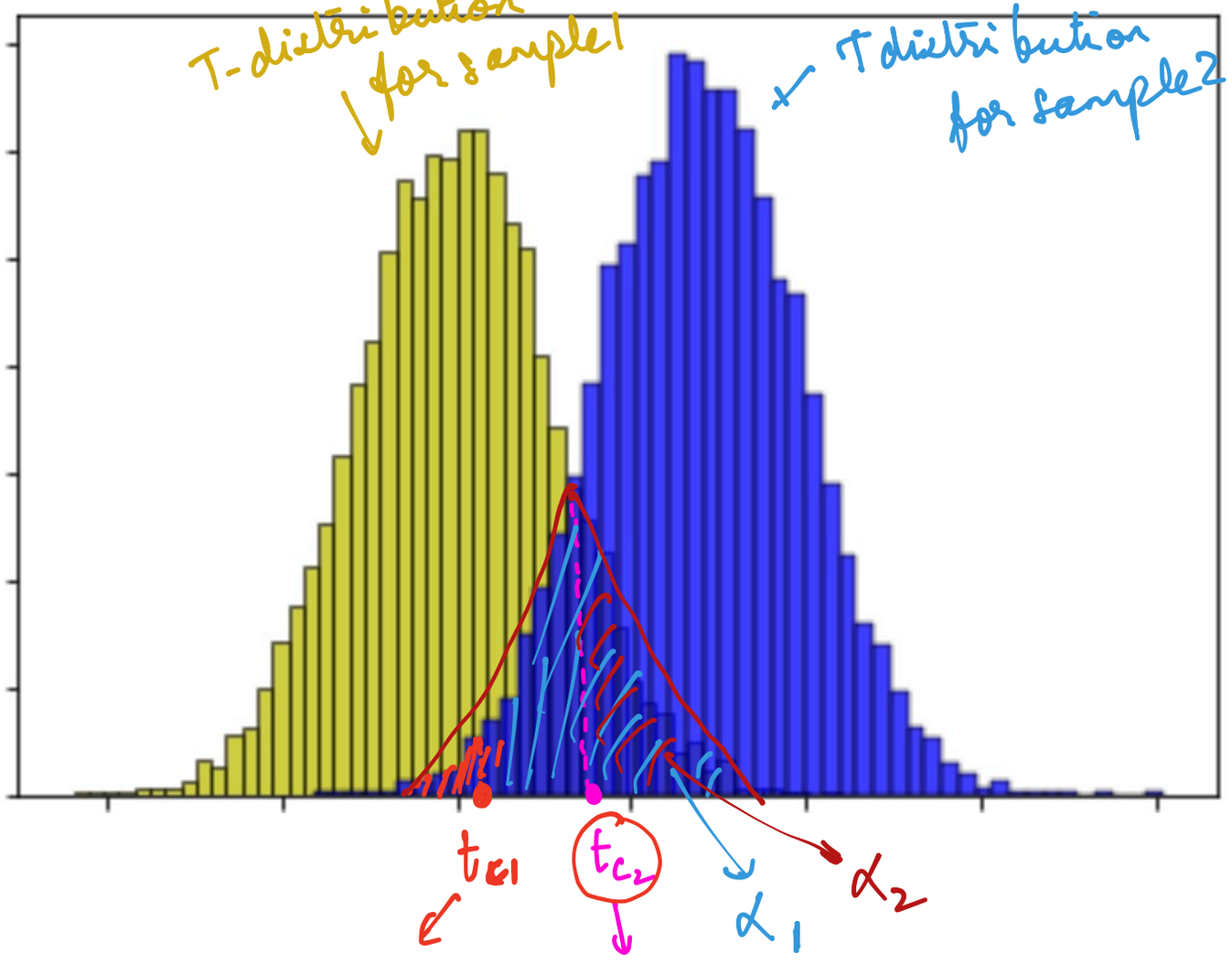
$$\left(\frac{t_c \times s \times \sqrt{2}}{E}\right)^2 = n$$

E → 0.1

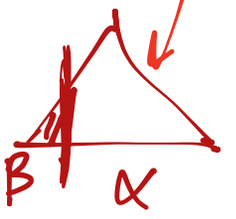
n that is just enough!

$$E = \frac{\mu_{difference}}{\sigma}$$

σ ↑ sample std.



n is controlling



$\rightarrow \beta_1 < \beta_2$.

$\rightarrow \alpha_1 > \alpha_2$

$\beta = 0.9$

$\alpha = 0.01$

S.E → defined for a sample.

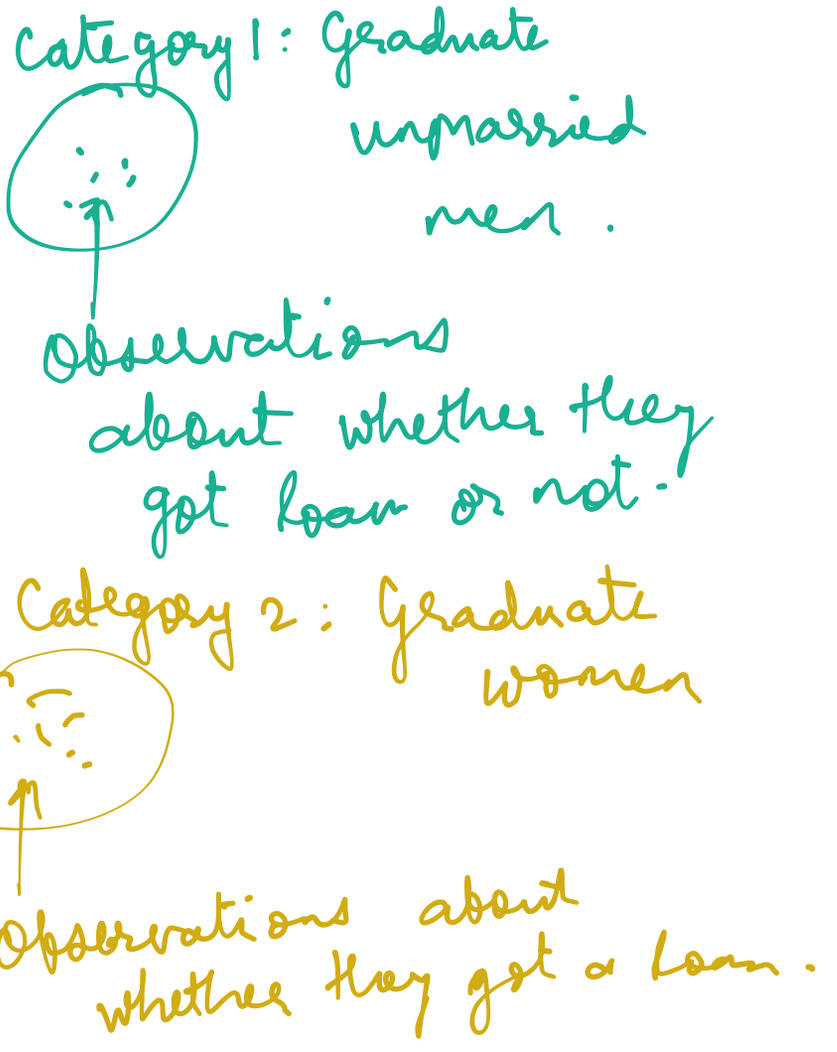
Sample $n = 7$

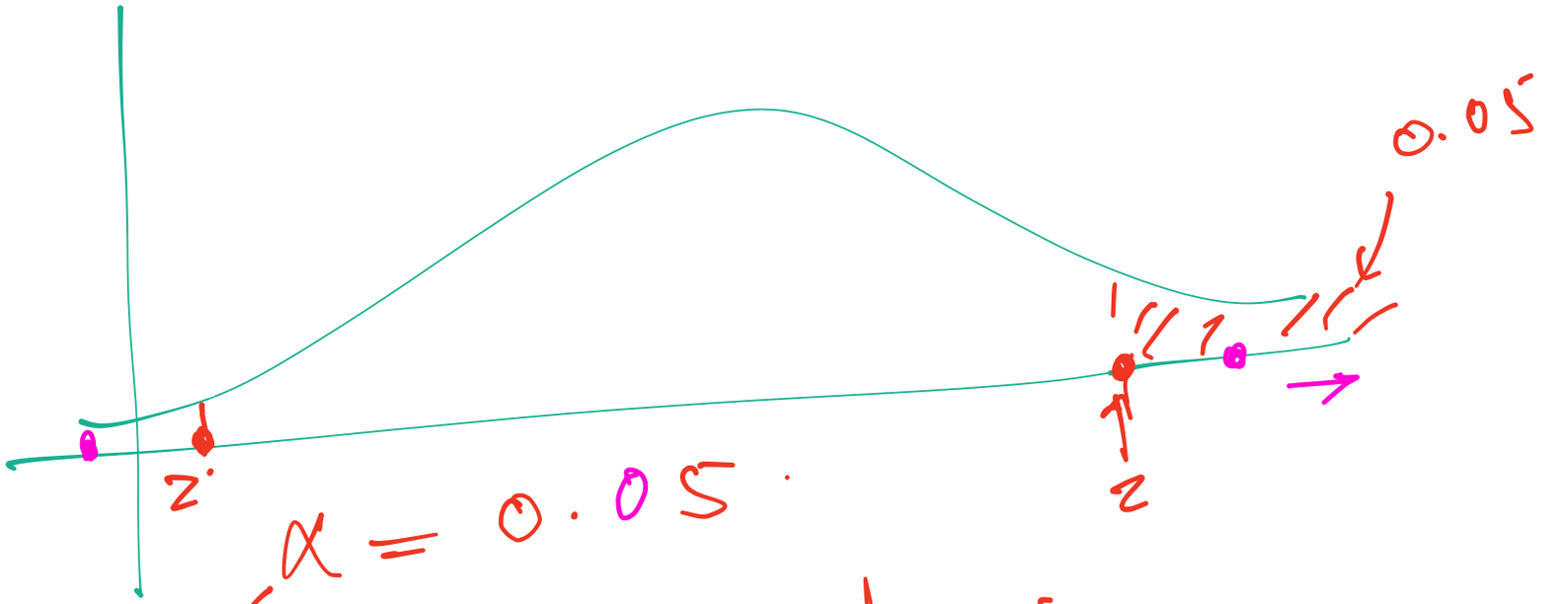
σ → Don't know.

Bootstrapping → Estimate " s "

$$S.E. = \frac{s}{\sqrt{n}}$$

We believe that the loan.csv data shows that graduate unmarried men are more likely to get a loan than graduate women. To prove this, would a t-test be more appropriate or a chi-square test? Carry out the appropriate test on the 'Loan_Status' column for the two groups and report the p-value. Also report your interpretation. Note: Assume a confidence level of 5% and round off the p-value to 2 decimal places.





$\alpha = 0.05$

Critical z-value =

(right-tailed) norm.ppf(0.95)

→ (left-tailed) → 0.05