Today's Agenda

1) Recap

2) Z-test — toy examples → scratch / scipy

3) T-Test — Case study → Youtube data

Giving a certain drug will increase IQ

H_0 : "drug is ineffective"

enough people

Population mean IQ = $\underline{\underline{\mu}}$ $\boxed{\hat{\mu}} \rightarrow$ IQ if you take drug

both of these we do not know

$H_0 : \tilde{\mu} = \mu$

$H_a : \tilde{\mu} \neq \mu$ $\left(H_a : \hat{\mu} > \mu \right)$

Population mean : IQ score "μ"   pop. std : σ

We make an assumption $H_0: \tilde{\mu} = \mu$

We get some samples / data → those who taken drugs (100 or 200)

We need an appropriate "test statistic" from the samples
Related to the sample mean → pop. mean

(1) We need the distribution T under H_0

(2) We need to compute p-value → accept Vs reject

$$\left(P\left[T \geq T_{obs}\right]\right)$$

T : "test statistic"

$H_0 : \tilde{\mu} = \mu$

This is the sample mean

$[x_1 \; x_2 \; x_3 \; \cdots \; x_n]$

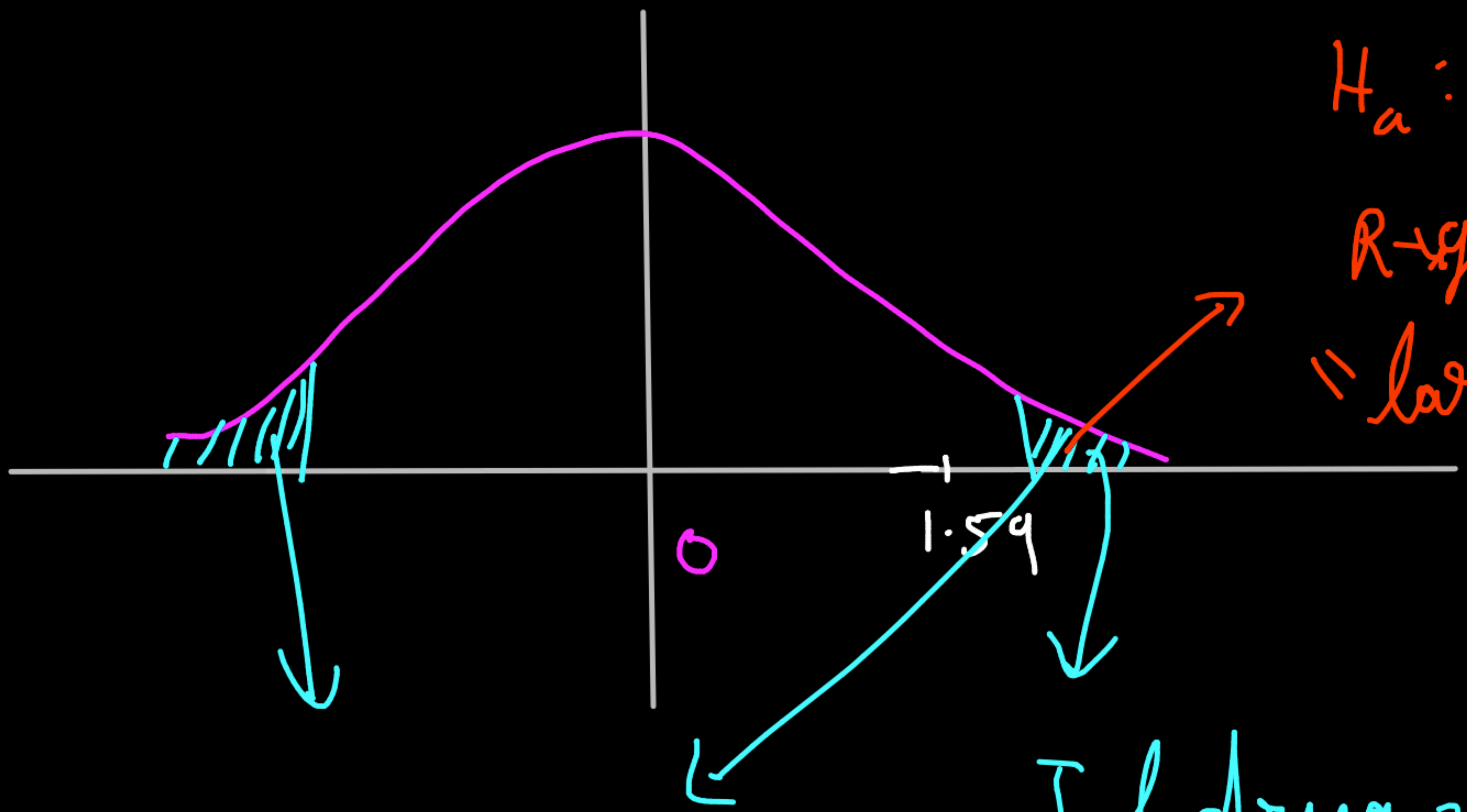
Sample mean $\quad m = \dfrac{x_1 + x_2 \cdots + x_n}{n} \sim N\left(\text{mean} = \mu, \; \text{std dev} = \dfrac{\sigma}{\sqrt{n}}\right)$

Standardization

$$\boxed{T = \frac{m - \mu}{(\sigma / \sqrt{n})}} \sim N(0, 1)$$

$$T = \frac{m - \mu}{\left(\frac{\sigma}{\sqrt{n}}\right)}$$

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



$H_a : \tilde{\mu} > \mu$

Right tail

"larger"

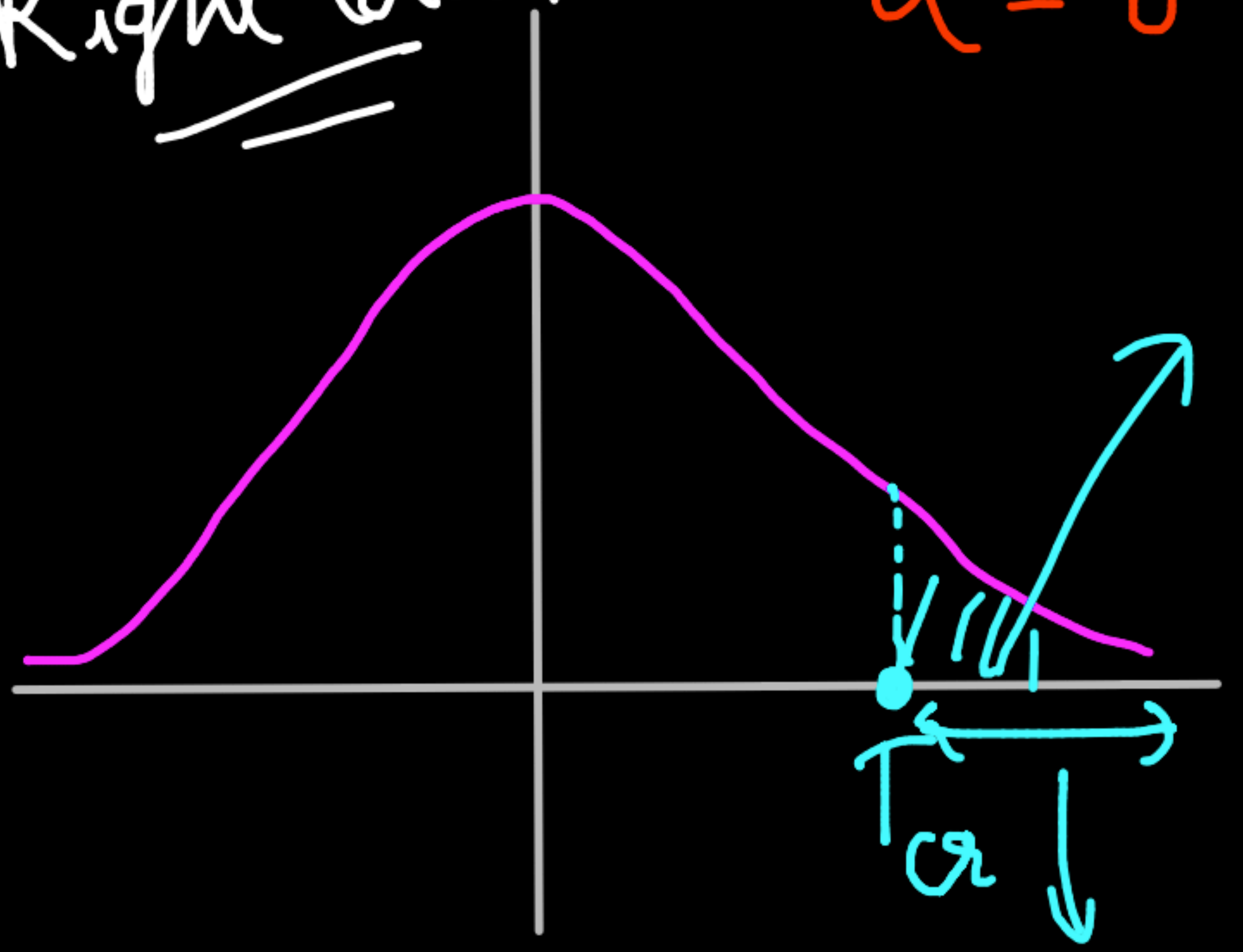
If drug worked observed statistic will lie in the right extreme

Right tailed

$\alpha = 0.05$

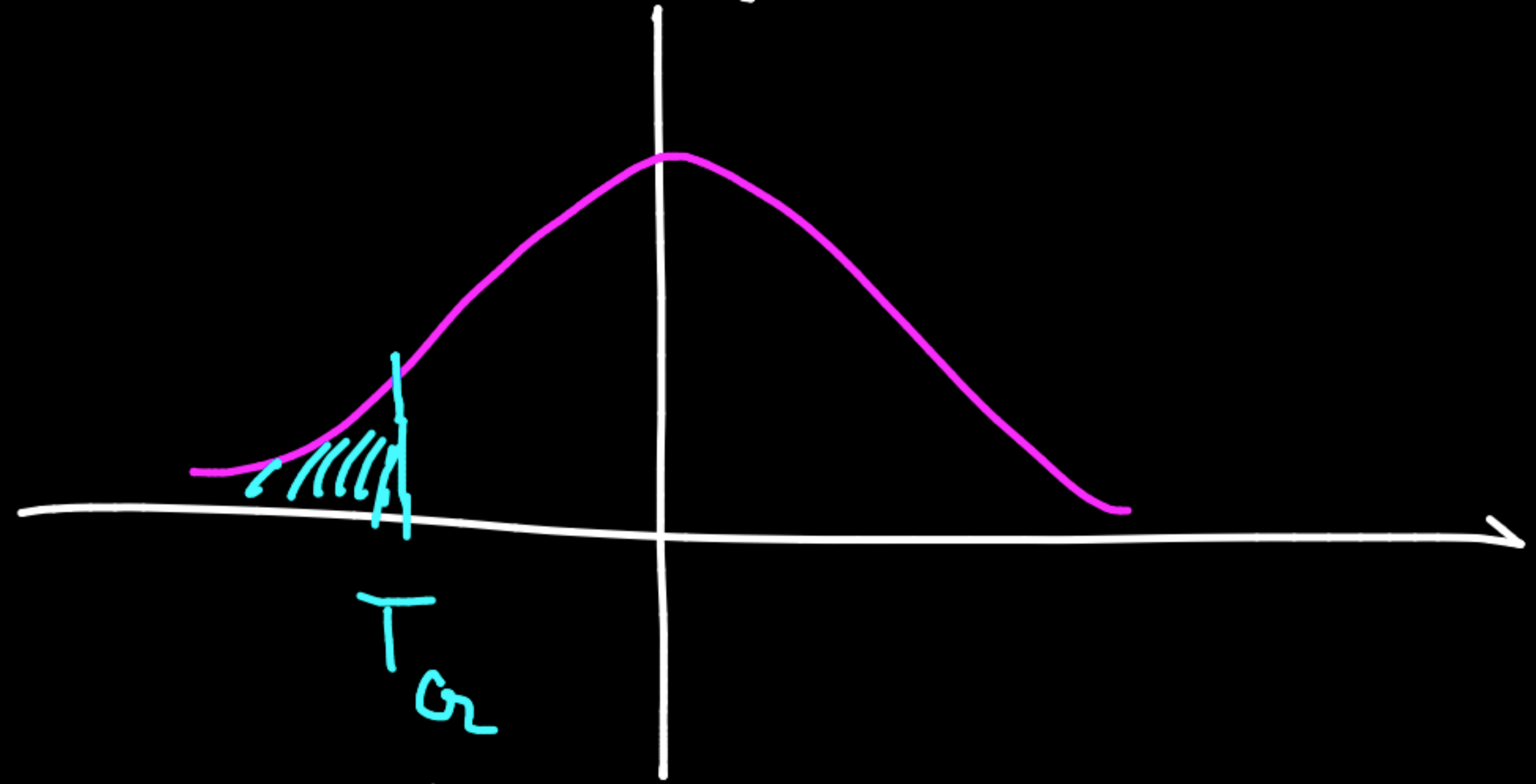
reject the H_0

$1 - cdf(0.05)$



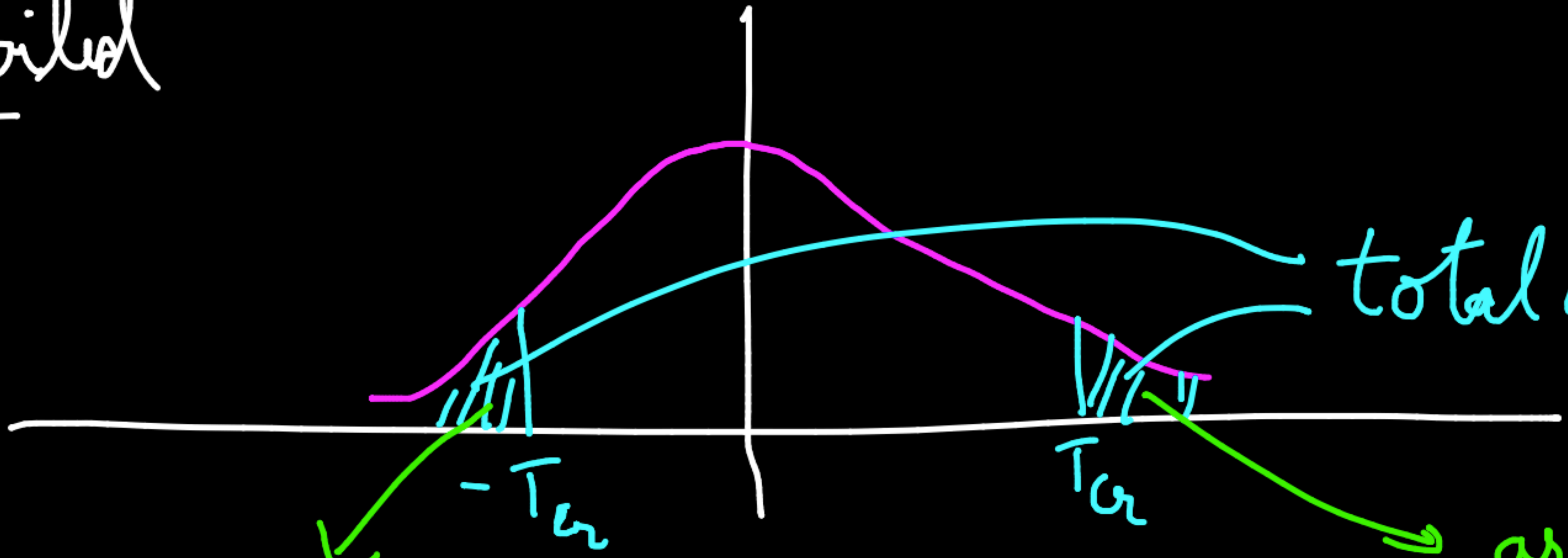
"critical region"

Left tailed



$cdf(0.05)$

Two-tailed



total area $= 0.05$

area $= 0.025$

0.025

Hypothesis test : We make a statement about the population using samples

Test statistic : Determined from the sample

Critical region : If "Observed test statistic" is in the critical region, we reject H_0

2 types of errors
→ Type 1 : H_0 is true, we reject → False positive
→ Type 2 : H_0 is false, we accept → False negative

Compare 2 drugs for recovery times : 2 sets of samples

Med 1 $\rightarrow$ recovery times $x_1 \; x_2 \; x_3 \cdots x_{n1}$ (say 150 patients)

Med 2 $\rightarrow$ recovery times $y_1 \; y_2 \; y_3 \cdots y_{n2}$ (say 200 people)

$\mu_1 \rightarrow$ pop. mean of med 1 $\qquad \mu_2 \rightarrow$ pop mean for med 2

$\sigma_1 \rightarrow$ pop std $\qquad \sigma_2 \rightarrow$ pop std dev

$$H_0 : \mu_1 = \mu_2$$

$$m_1 = \frac{x_1 + x_2 + \cdots x_{n1}}{n_1} \qquad m_2 = \frac{y_1 + y_2 + \cdots y_{n_2}}{n_2}$$

Options:

$$m_1 - m_2$$

$$m_1 / m_2$$

$$m_1 - m_2 / S.E$$

1) Med 1 $\mu_1, \sigma_1 \rightarrow$ population paramete

$n_1, m_1, s_1 \rightarrow$ no. / sample mean / sample stdv

2) Med $\mu_2 \; \sigma_2 \rightarrow$ pop

$n_2 \; m_2 \; s_2 \rightarrow$ sample

$H_0 : \mu_1 = \mu_2$

$H_a : \mu_1 \neq \mu_2$

$$m_1 \sim N\left(\mu_1, \frac{\sigma_1}{\sqrt{n_1}}\right) \qquad E[m_1] = \mu_1 \qquad Var[m_1] = \frac{\sigma_1^2}{n_1}$$

$$m_2 \sim N\left(\mu_2, \frac{\sigma_2}{\sqrt{n_2}}\right) \qquad E[m_2] = \mu_2 \qquad Var[m_2] = \frac{\sigma_2^2}{n_2}$$

Guess : $m_1 - m_2$

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

$$Var[m_1 - m_2] = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

Standardize

$$\frac{(m_1 - m_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \sim N(0, 1)$$

$$X \longrightarrow \mathrm{Var}[X] = \sigma^2$$

$$\mathrm{Var}[aX] = a^2 \mathrm{Var}(X) = a^2 \sigma^2$$

$$\mathrm{Var}[-X] = (-1)^2 \mathrm{Var}(X) = \sigma^2$$

$$\mathrm{Var}[X+Y] = \mathrm{Var}(X) + \mathrm{Var}[Y] \quad \text{(if independent)}$$

$$\mathrm{Var}(m_1 - m_2] = \mathrm{Var}\, m_1 + \mathrm{Var}(m_2]$$

Safari File Edit View History Bookmarks Window Help 100% Mon 25 Jul 11:03 PM

Summary

1) One set of samples Vs population parameter

$x_1 \; x_2 \cdots x_n$

$\mu = 100$ (IQ)

2) two sets of samples

$x_1 \; x_2 \cdots x_{n_1}$

$y_1 \; y_2 \cdots y_{n_2}$

Safari File Edit View History Bookmarks Window Help   100%   Mon 25 Jul 11:35 PM

Test-Statistic :

One-sample $\quad \dfrac{m - \mu}{(\sigma)/\sqrt{n}} \quad H_0 : \tilde{\mu} = \mu$

Two sample $\quad \dfrac{m_1 - m_2}{\sqrt{\dfrac{(\sigma_1)^2}{n_1} + \dfrac{(\sigma_2)^2}{n_2}}} \quad H_0 : \mu_1 = \mu_2$

(S) instead $\quad n < 30 \rightarrow$ t-test

$n > 30 \rightarrow$ z or t

2 possibilities regarding σ

we know σ → z-test

we do not know σ → n is small, $n < 30$ → $\boxed{\text{t-test}}$

we do not know σ → n is large, $n > 30$ → both t & z test work

S instead of σ

S → Sample std dev

$$T = \frac{m - \mu}{\sqrt{s/\sqrt{n}}}$$

s : sample std dev

no longer Gaussian

"t- distribution"

degrees of freedom $n-1$

As $n \to \infty$, $t \longrightarrow$ Gaussian

$$T = \frac{m - \mu}{\sqrt{s/\sqrt{n}}}$$

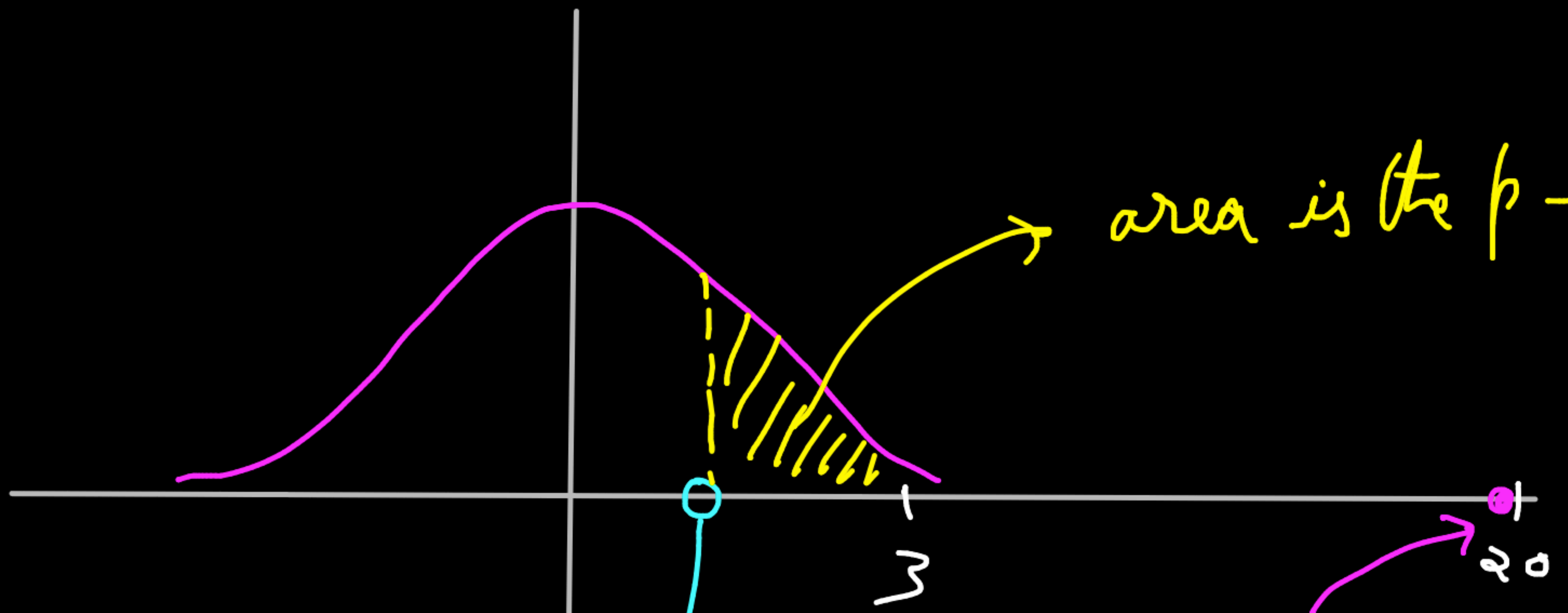
s : sample std dev

no longer Gaussian

" t- distribution "

degrees of freedom $n-1$

As $n \to \infty$, $t \longrightarrow$ Gaussian



area is the p-value

$P[T \geq T_{obs}]$

$P[T \geq 1.5]$

$T_{obs} = 1.5$

"value" observed test statistic

100 - 110,

standari

If test statistic is Standard Gaussian

"Z score"