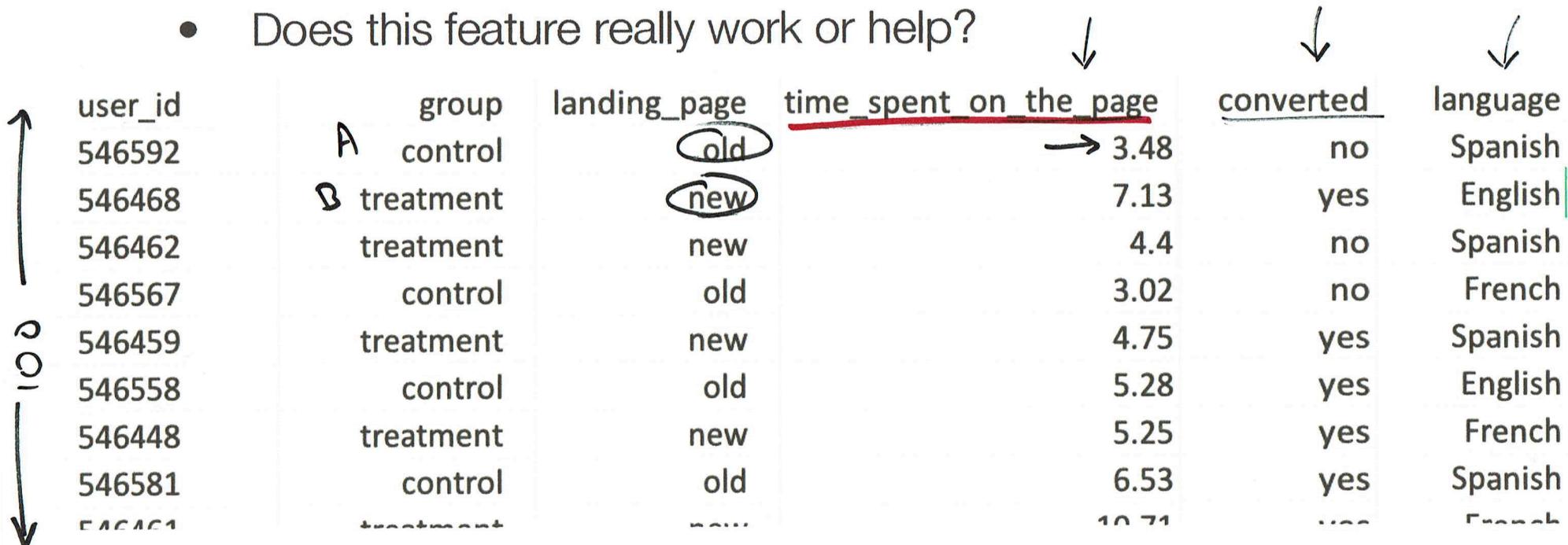


Lecture plan

- Fundamentals of Statistics
 - Summarizing Data
 - Understanding and modeling Random Variables
 - The Issue: Unknown 'Population' Vs known 'Sample'
 - The Idea: Central Limit Theorem (CLT)
- Hypothesis testing
 - A/B Testing
 - Other tests
- Anomaly Detection

A/B Testing

- Online news portal aims to expand its business by acquiring new subscribers.
- A team analyzes user behaviors and comes up with a new feature.
- Does this feature really work or help?



The diagram shows a table of user data with two vertical arrows on the left. The top arrow points upwards and is labeled 'A', indicating the control group. The bottom arrow points downwards and is labeled 'B', indicating the treatment group. The table has columns for user_id, group, landing_page, time_spent_on_the_page, converted, and language. The 'time_spent_on_the_page' column is underlined in red. The 'converted' column has a red arrow pointing to 'no' for the first row and 'yes' for the second row. The 'language' column has a red arrow pointing to 'Spanish' for the first four rows and 'French' for the last three rows. The 'group' column has a red arrow pointing to 'control' for the first four rows and 'treatment' for the last three rows. The 'landing_page' column has a red arrow pointing to 'old' for the first two rows and 'new' for the last two rows.

user_id	group	landing_page	time_spent_on_the_page	converted	language
546592	A control	old	3.48	no	Spanish
546468	B treatment	new	7.13	yes	English
546462	treatment	new	4.4	no	Spanish
546567	control	old	3.02	no	French
546459	treatment	new	4.75	yes	Spanish
546558	control	old	5.28	yes	English
546448	treatment	new	5.25	yes	French
546581	control	old	6.53	yes	Spanish
546461	treatment	new	10.71	yes	French

- Why is this a difficult problem?

A → 4.5
B → 6.2

(A)

50

(31) Heads

~~22~~

(B)

50

(22) Heads

~~21~~

$x_1, x_2, \dots, \dots, \dots, x_{100}$

○ Measure of center \rightarrow Mean/Avg., Median, Mode

$$\sum x_i / 100 = \bar{x}$$

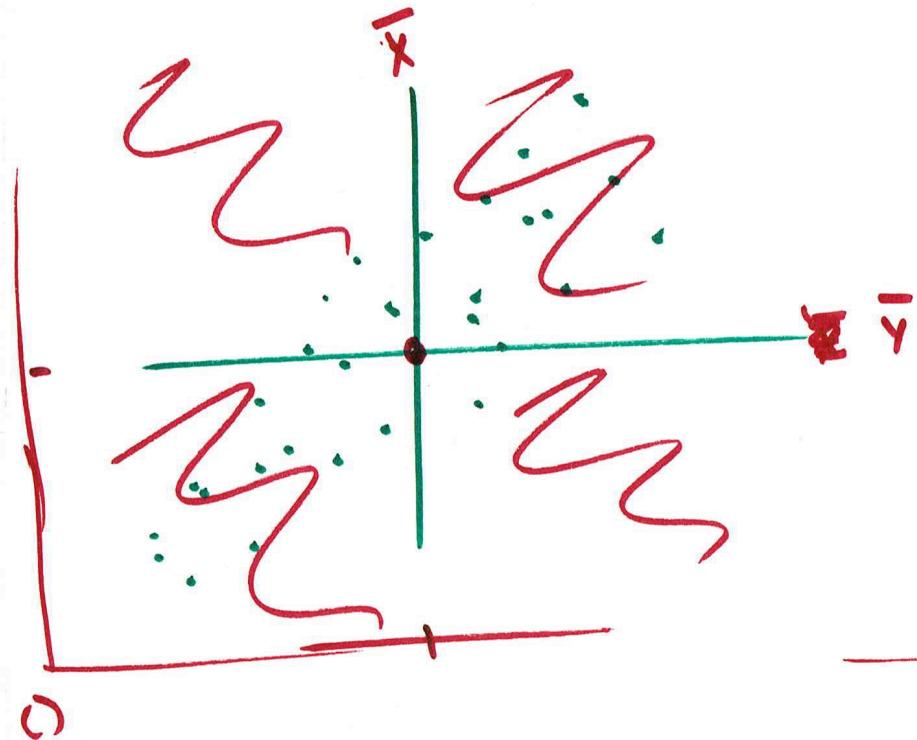
○ Measure of spread \rightarrow ~~Range = $\max(x_i) - \min(x_i)$~~

~~Any dist from center } = $\sum |x_i - \bar{x}| / 100$~~

Risk \Rightarrow $\boxed{\text{Var} = \frac{\sum (x_i - \bar{x})^2}{100}}$

$\text{Std dev. } \Rightarrow \sqrt{\frac{\sum (x_i - \bar{x})^2}{100}}$

$x_1, \dots, \dots, \dots, x_{100}$ } ←
 $y_1, \dots, \dots, \dots, y_{100}$



$$\text{Covar} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$\text{Cov}(x, y) = \text{Cov}(y, x)$$

$$\text{Cov}(x, x) = \text{Var}(x)$$

$$\text{Corr}(x, y) = \frac{\text{Cov}(x, y)}{\text{Std}(x) \cdot \text{Std}(y)} = \frac{0.8}{\frac{1}{\sqrt{2}}} = 0$$

↑ +1
↓ -1

Deterministic Var

$$4x + 10 = 90$$

$$\boxed{x = 20} \quad \boxed{\text{Value}}$$

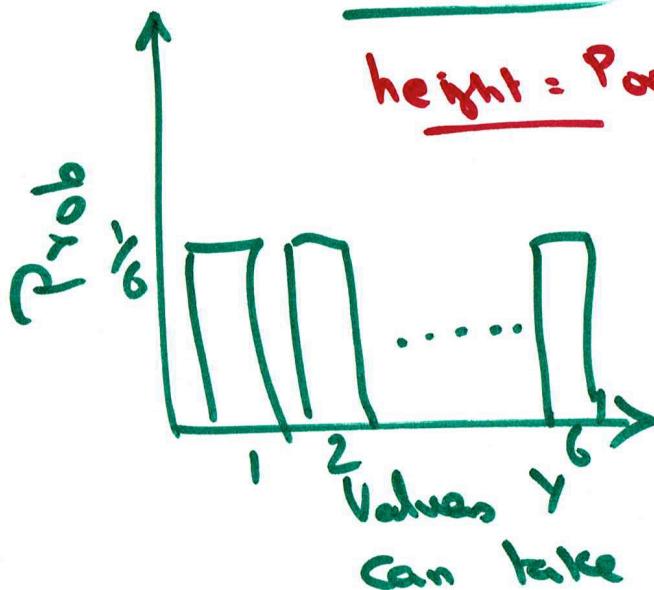
Random Var

$Y = \begin{cases} \textcircled{1} \text{ the values } Y \text{ can take} \\ \textcircled{2} \text{ the corresponding probabilities} \end{cases}$

Distribution

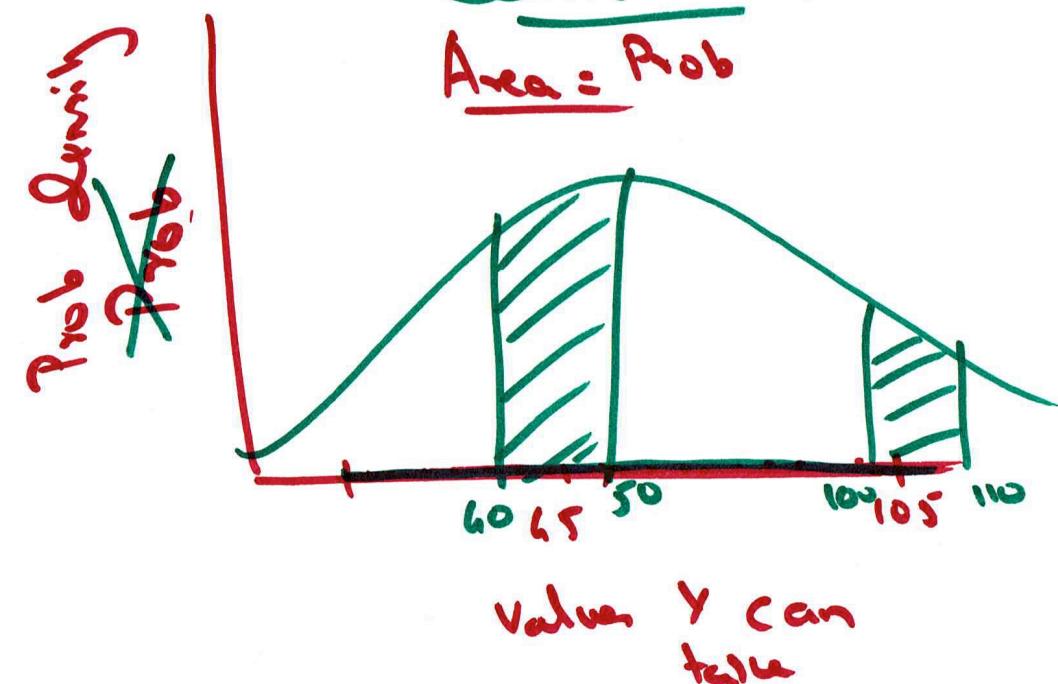
Discrete R.V.

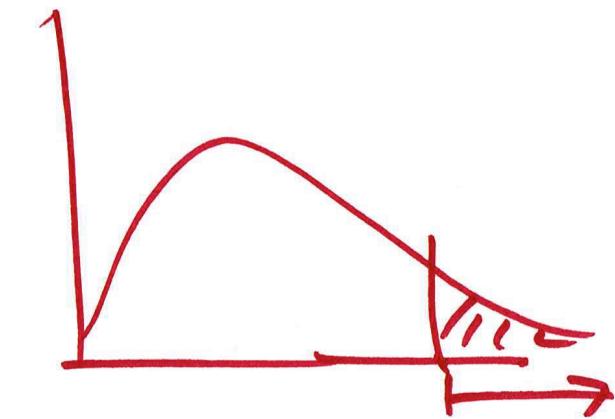
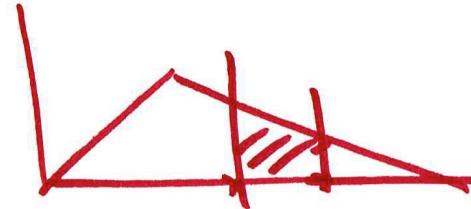
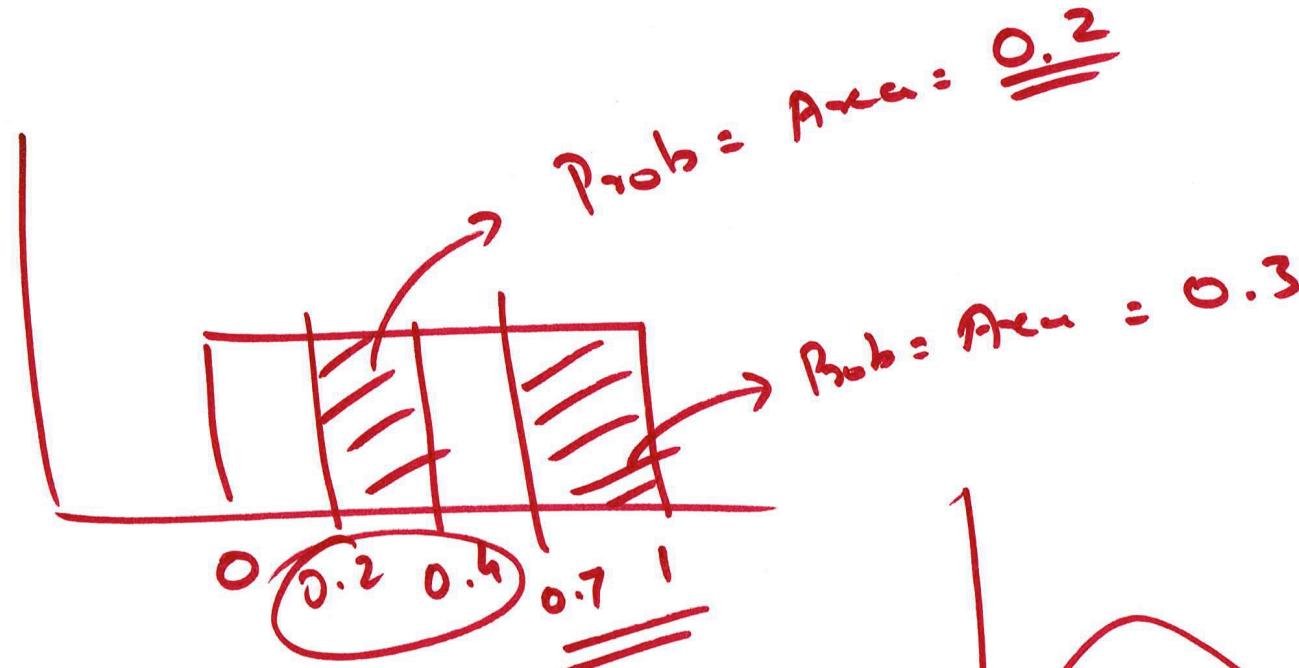
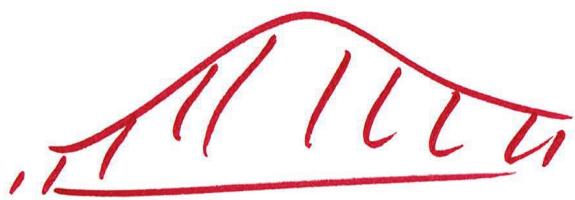
height = Prob



Cont. R.V.

Area = Prob

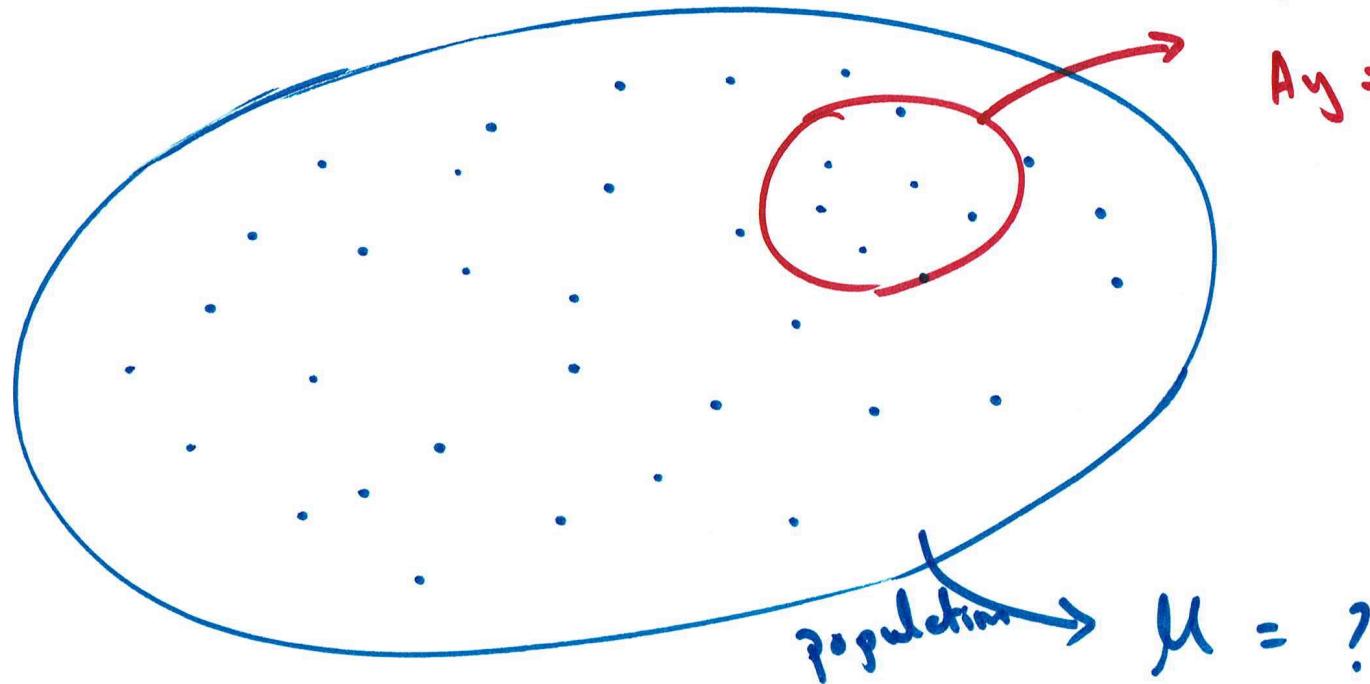




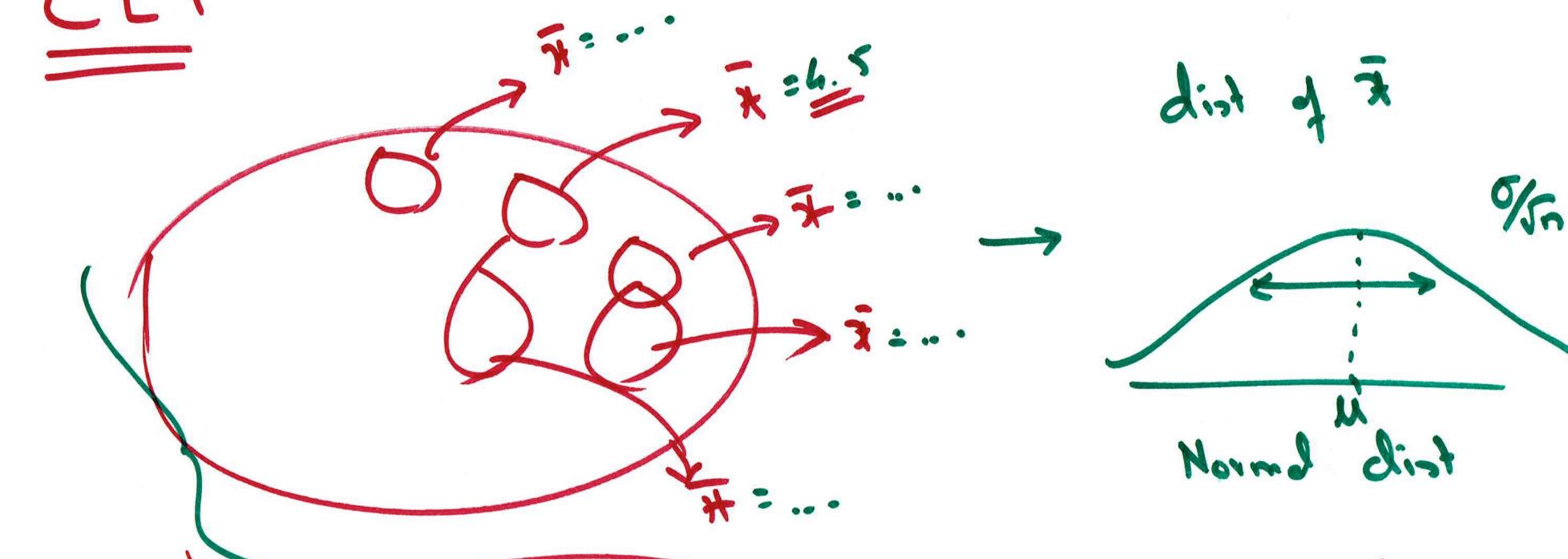
Any amt. of time people spend
on our website

Sample
 $n = 50$

$\bar{A}y = 4.5 \text{ min}$



CLT

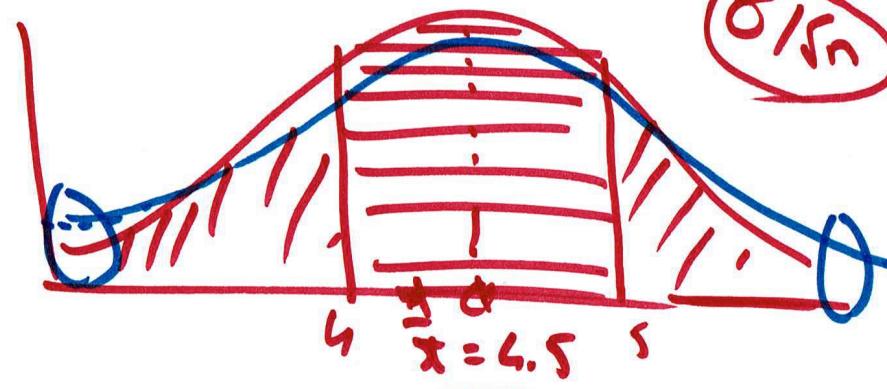


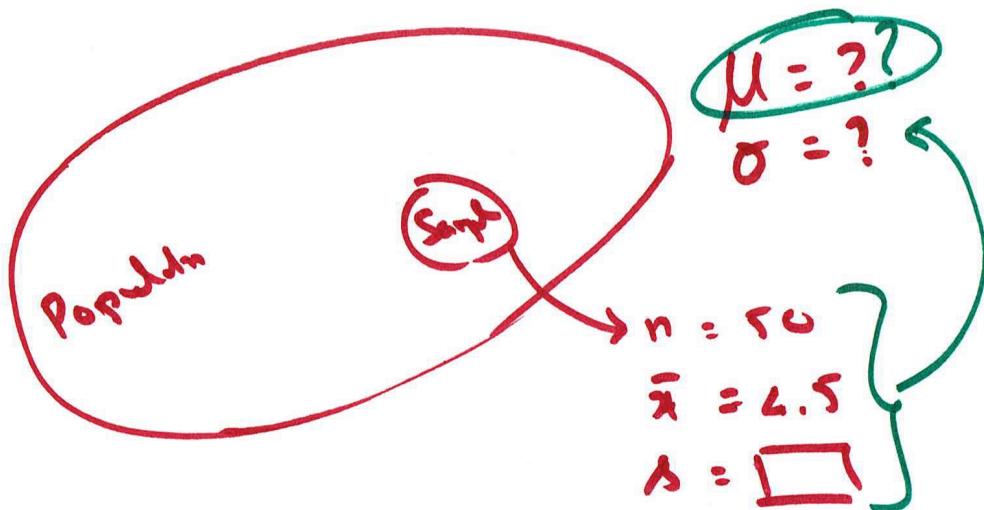
true avg
 μ

dist. of unknown μ

Normal mean
 (σ/\sqrt{n})

std. dev.
 σ/\sqrt{n}

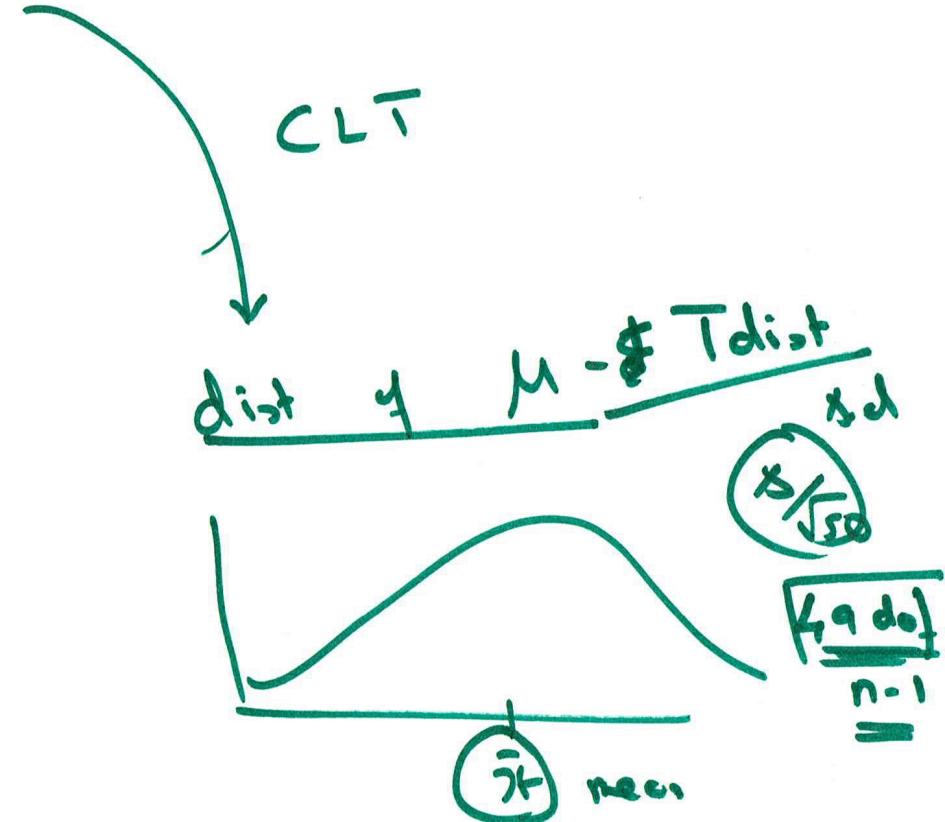




4 mins

Hypothesis Test

Null Hypothesis $H_0: \mu = 4$
 $H_a: \mu > 4$



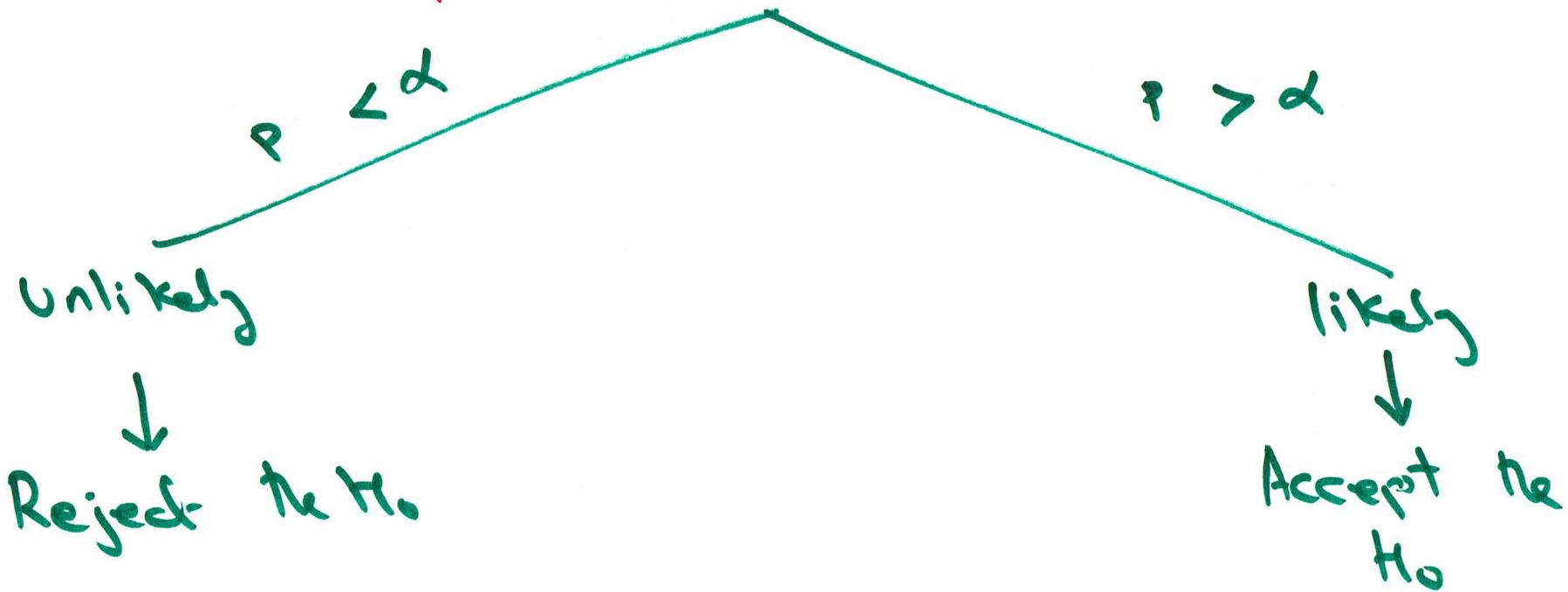
Assume : $H_0: \mu = 4$

$H_a: \mu > 4$

Evidence : $\bar{x} = 4.5$ $n = 50$

.....

Question : what are the chances of observing
the ~~evidence~~ an \bar{x} of 4.5 or more
if our $\boxed{\mu = 4}$ assumption are true



$$\alpha = 5.1 \cdot 10^{-11}$$

