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?

user_id	group	landing_page	time_spent_on_the_page	converted	language
546592	control	old	3.48	no	Spanish
546468	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

- Why is this a difficult problem?

A → 4.5
B → 6.2

A

50

31 Heads

~~22~~

B

50

22 Heads

~~28~~

$x_1 \quad x_2 \quad \ldots \ldots \ldots \quad x_{100}$

⊙ Measure of Center → Mean/Avg, Median, Mode

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

⊙ Measure of Spread (Risk) → ~~Range = max(x_i) - min(x_i)~~

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

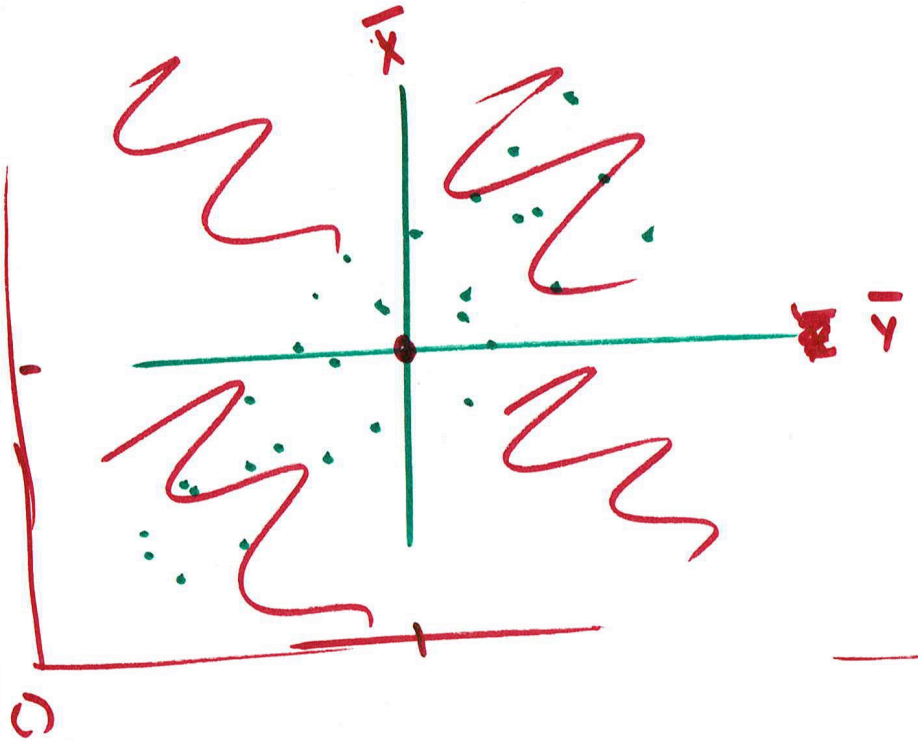
Risk ⇒ $\boxed{\text{Var} = \frac{\sum (x_i - \bar{x})^2}{100}}$

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

?

$X_1 \ldots\ldots\ldots X_{100}$

$Y_1 \ldots\ldots\ldots Y_{100}$ ←



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

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

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

$$\text{Cor}(x, y) = \frac{\text{Cov}(x, y)}{\text{Std}(x)\,\text{Std}(y)} = 0.8$$

+1

0

−1

Deterministic Var

$4x + 10 = 90$

$x = 20$ ← Value

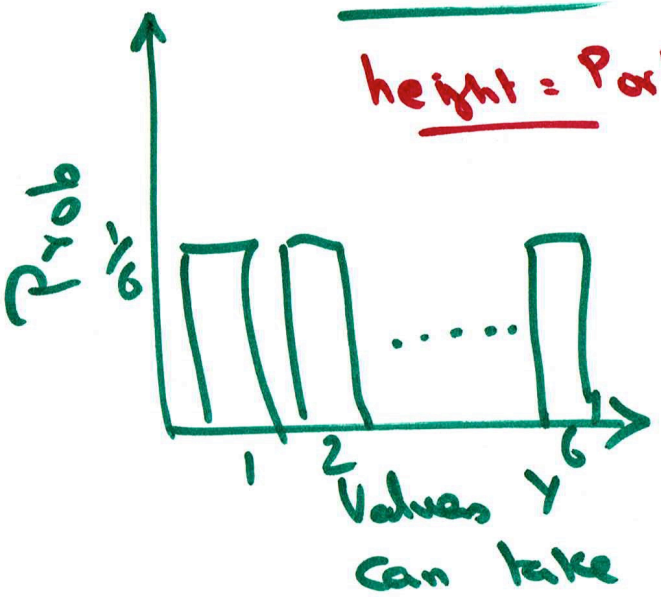
Random Var

$Y =$ { a) the values Y can take
b) the corresponding probabilities

→ Distribution

→ Discrete R.V.

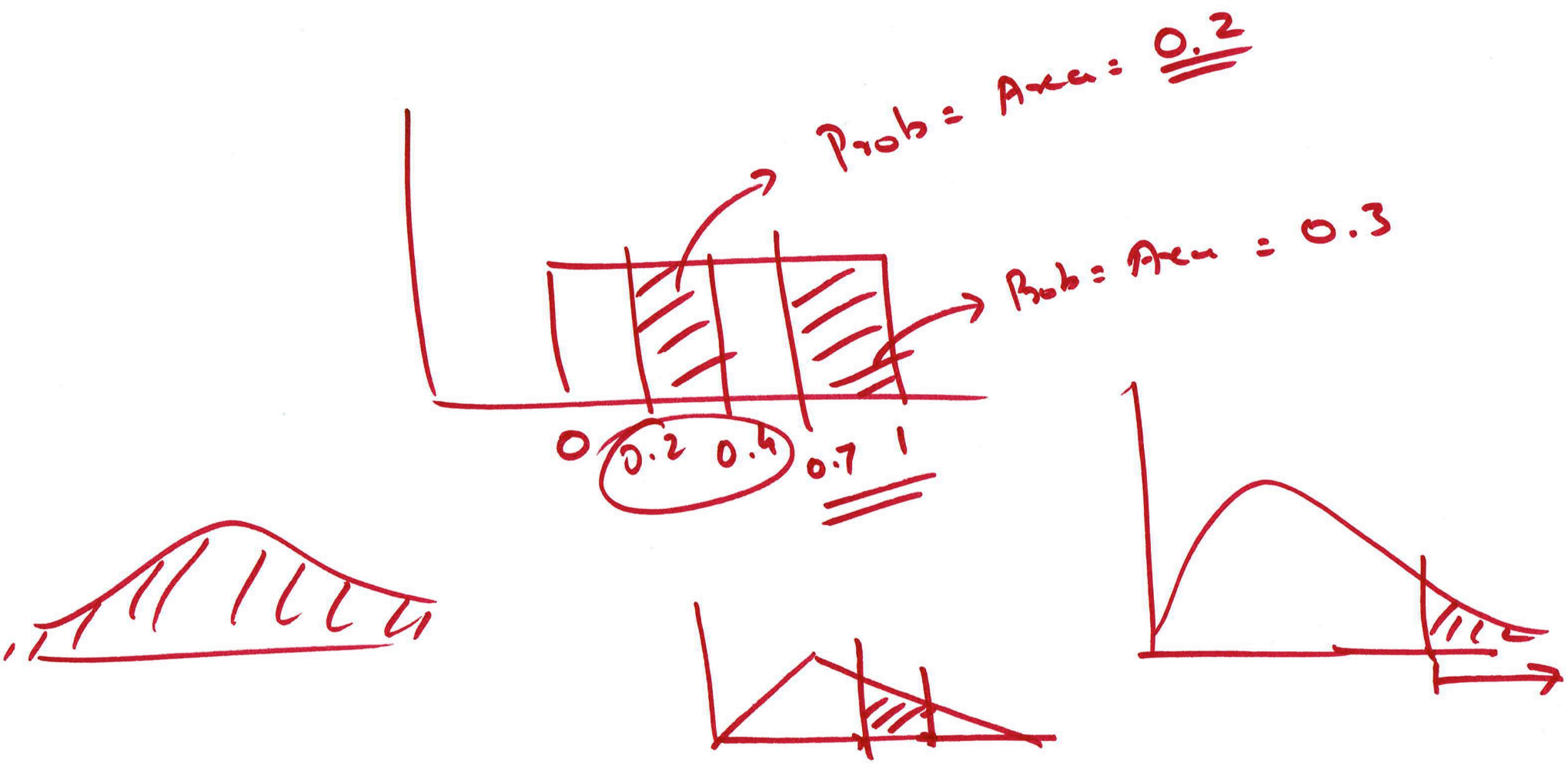
height = Prob



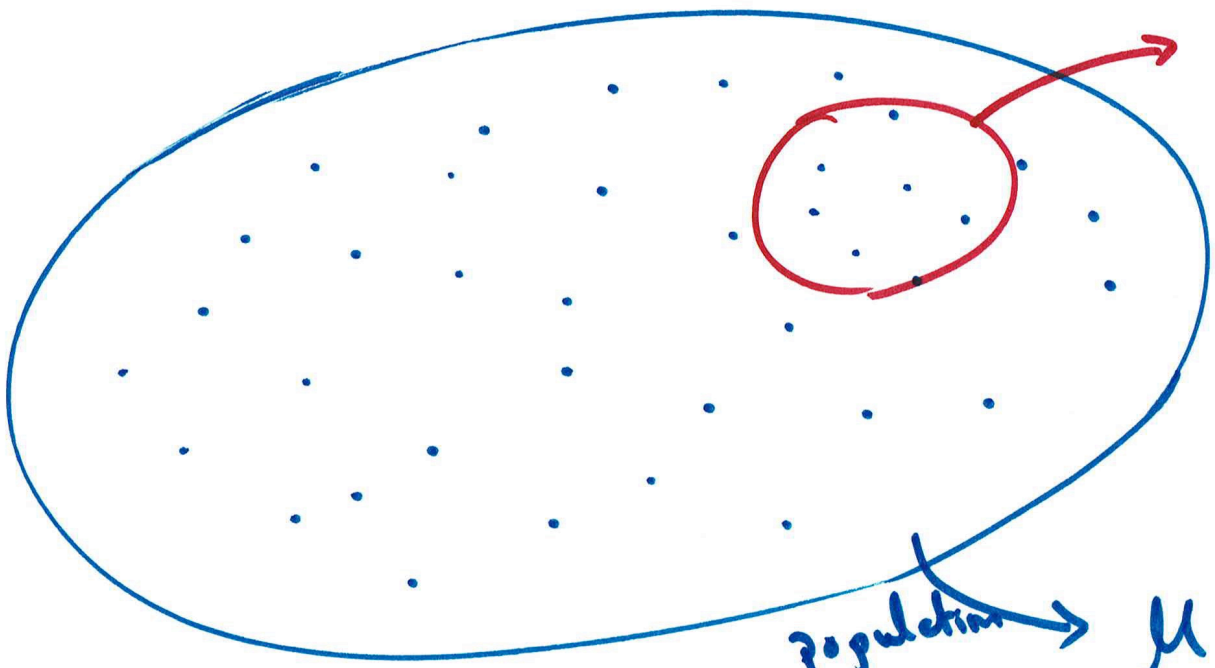
→ Cont. R.V.

Area = Prob

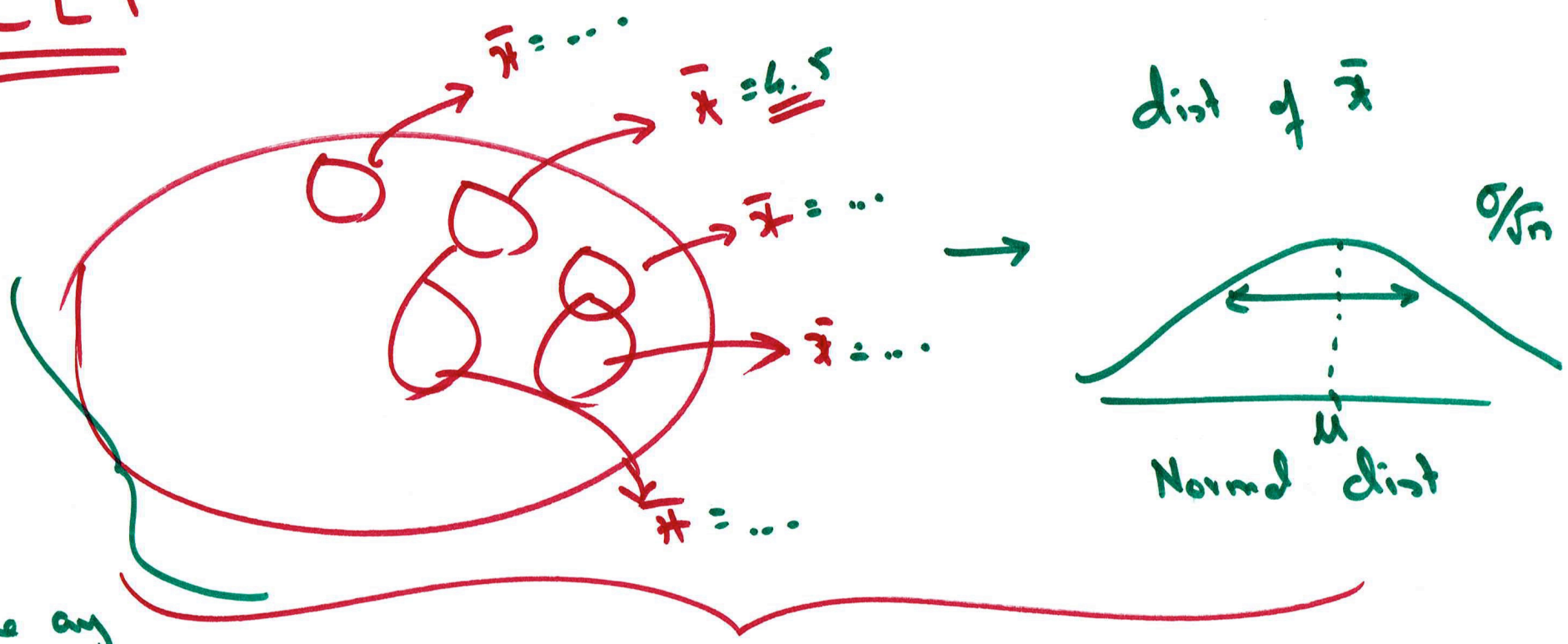
Prob ~~Prob~~


Prob = Area = 0.2
Prob = Area = 0.3
0
0.2
0.4
0.7
1

Ay amt. of time people spend on our website



CLT

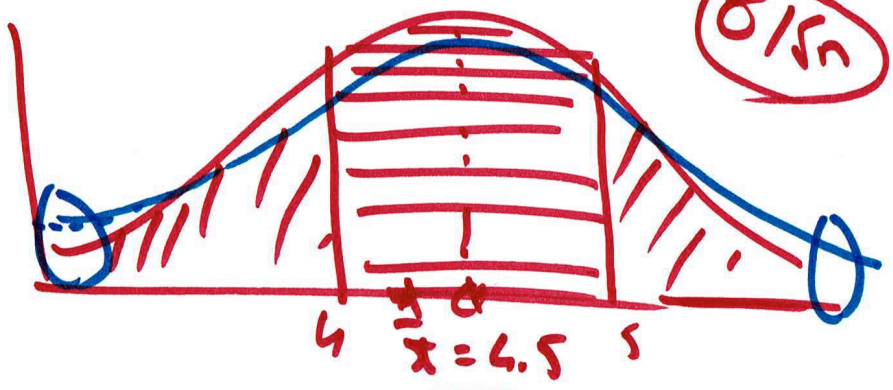


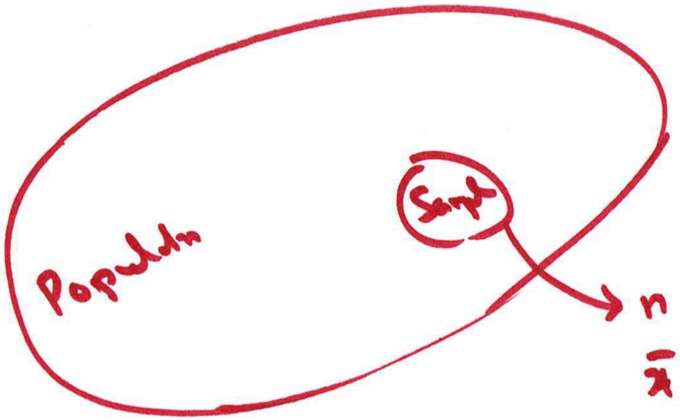
true avg μ

dist. of unknow μ

Normal
mean ✓
$\sigma/\sqrt{n}$ ✓

T dist
~~the~~ mean ✓
$s/\sqrt{n}$ ✓
dof ✓





$\mu = ??$

$\sigma = ?$

$n = 50$

$\bar{x} = 4.5$

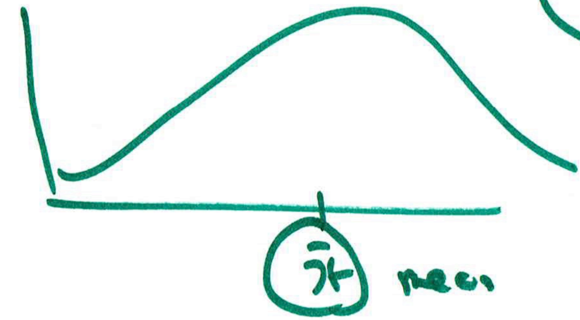
$s = [\ ]$

CLT

dist of μ - T dist

sd

$s/\sqrt{50}$



49 df

n-1

4 mins

Hypothesis Test

Null Hyp $H_0: \mu = 4$

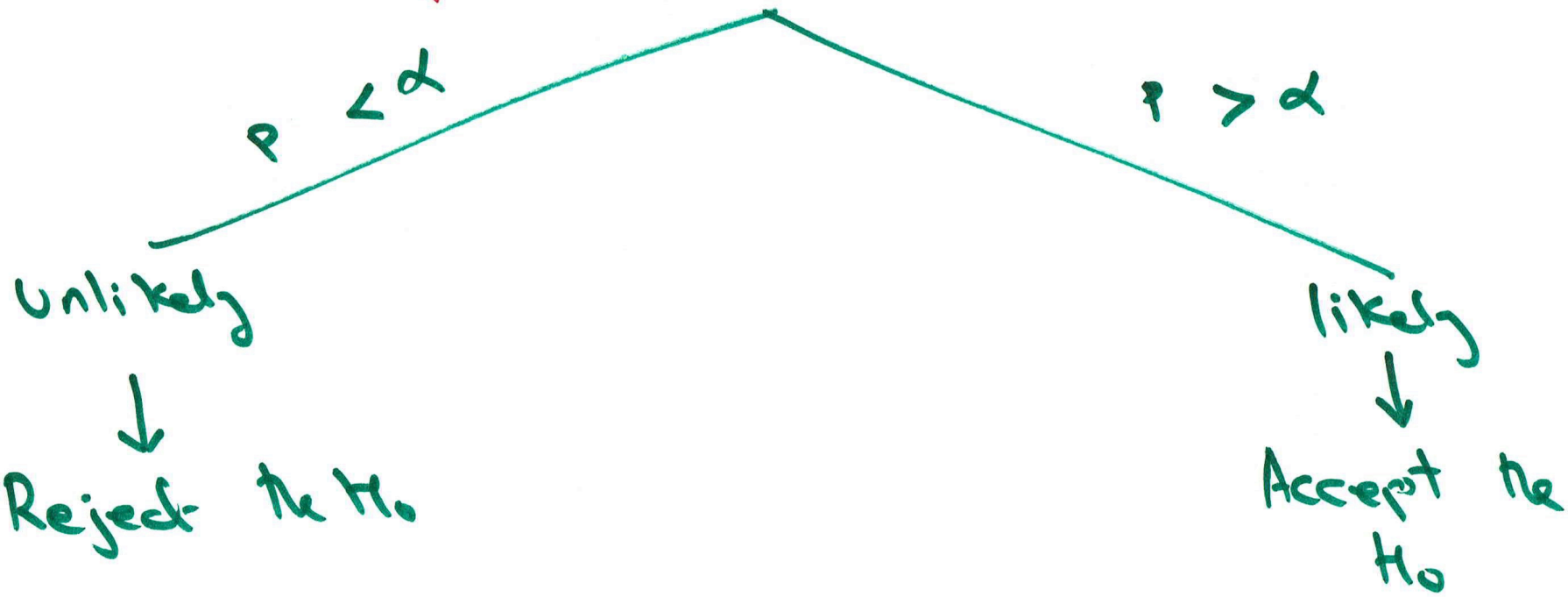
$H_a: \mu > 4$

Assume: $H_0: \mu = 4$

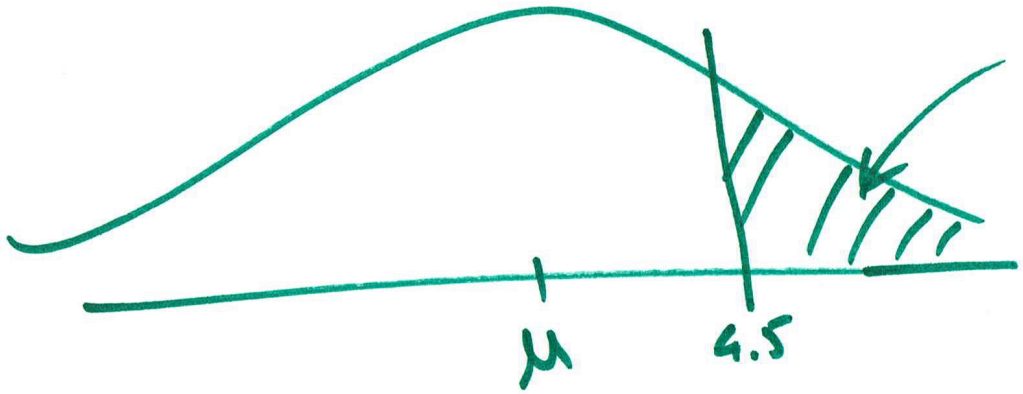
$H_a: \mu > 4$

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

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



$\alpha = 5\%$ or 10%



Prob = 0.4 ~~0.004~~