# Inferencial statistics

- Null vs Alternative hyp

## Steps in data-driven decision making.

1. Formulate a hypothesis
2. Find the right test
3. Execute the test
4. Make a decision

Hyp := "An idea that can be tested"

**Example** "Glassdoor data science salaries"

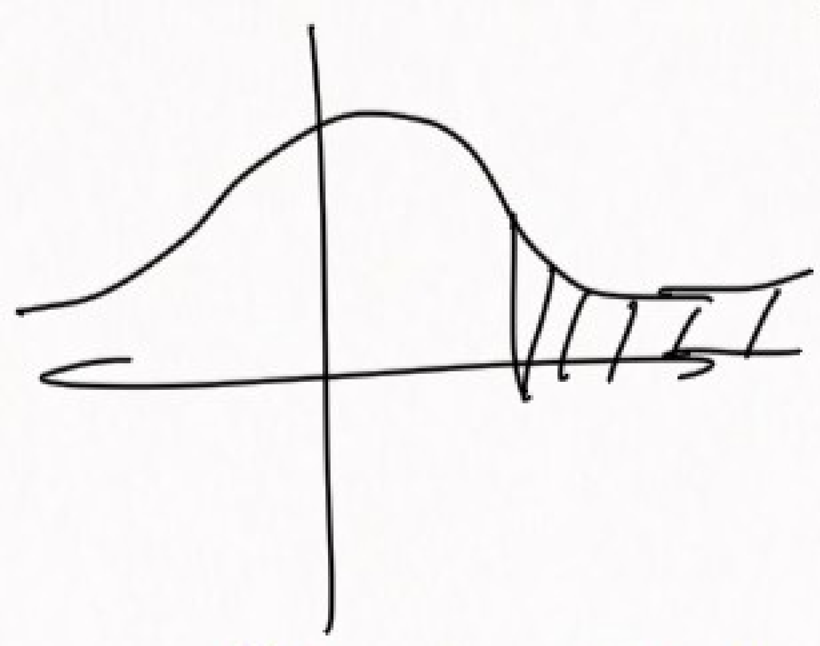
$\mu = \$113k$ USD. (According to Glassdoor this is the avg salary for data scientist is USA.)

$$\begin{cases} H_0 := \mu = \$113.000 \\ H_1 := \mu \neq \$113.000 \end{cases}$$

Two sided test: if $\mu \neq 113{,}000$

$\Rightarrow$ Either $\mu > 113{,}000$ or $\mu < 113.000$

# One sided tests



$H_0 := \mu_0 \geq \$ 125,000$

$H_1 := \mu_0 < \$ 125.000.$

• Chances of making a mistake: "Reject Ho when its true".

$\alpha$ := probability of rejecting Ho, if its true

Tipical values of $\alpha$ 0,01, 0,05 or 0,1

Example "University avg grade."

You work as a consultant for a public university and want to carry out an analysis on how students are performing on avg.

The dean believes that on avg students have a GPA of 70%.

LET'S TEST:

$H_0 := \mu_0 = 70\%$

$H_1 := \mu_0 \neq 70\%$

Assuming NORMAL DISTRIBUTION of the students grades.

# Z - test

sample mean

hypothesized mean

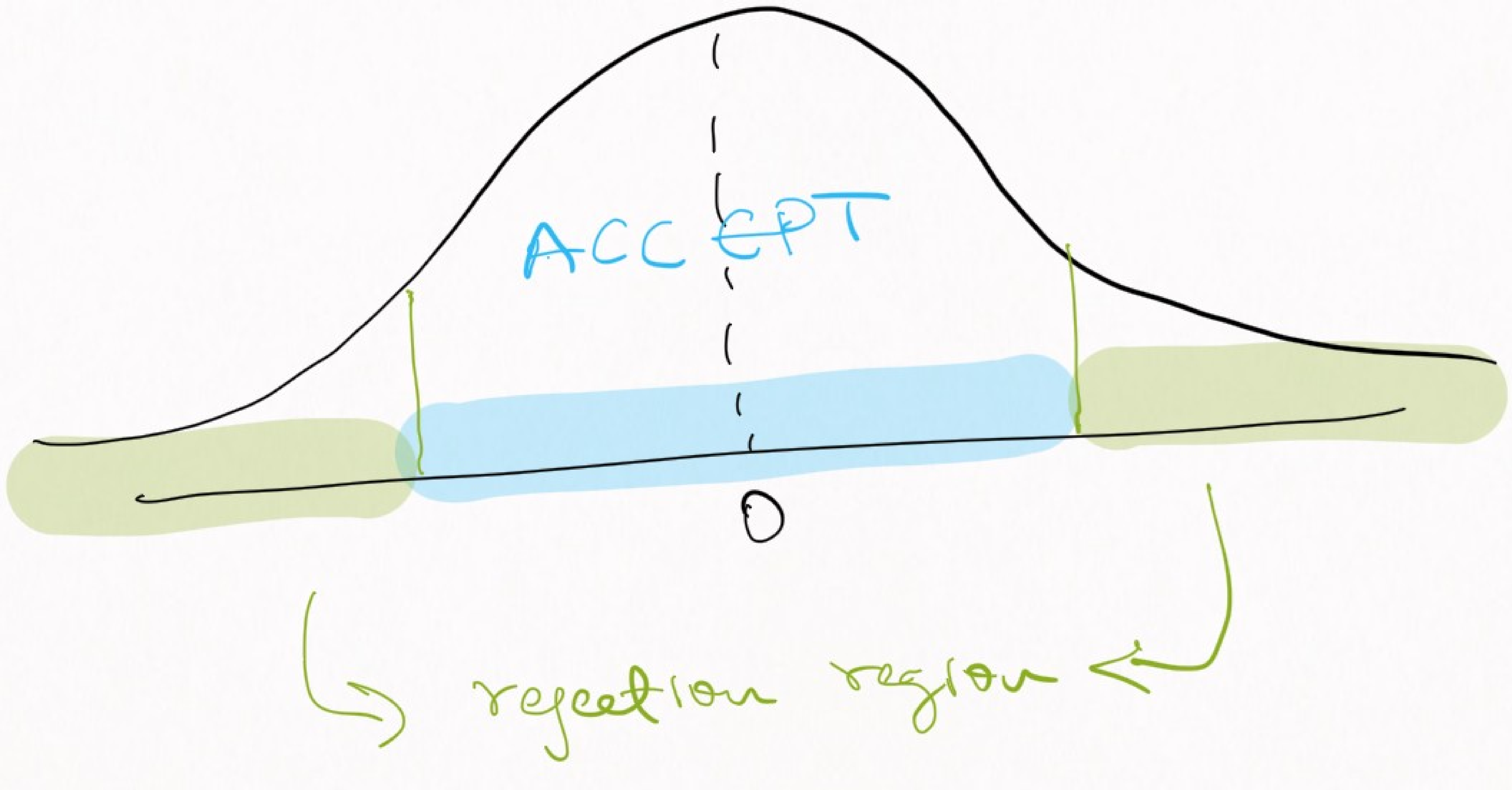
$$Z = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

↳ standard error

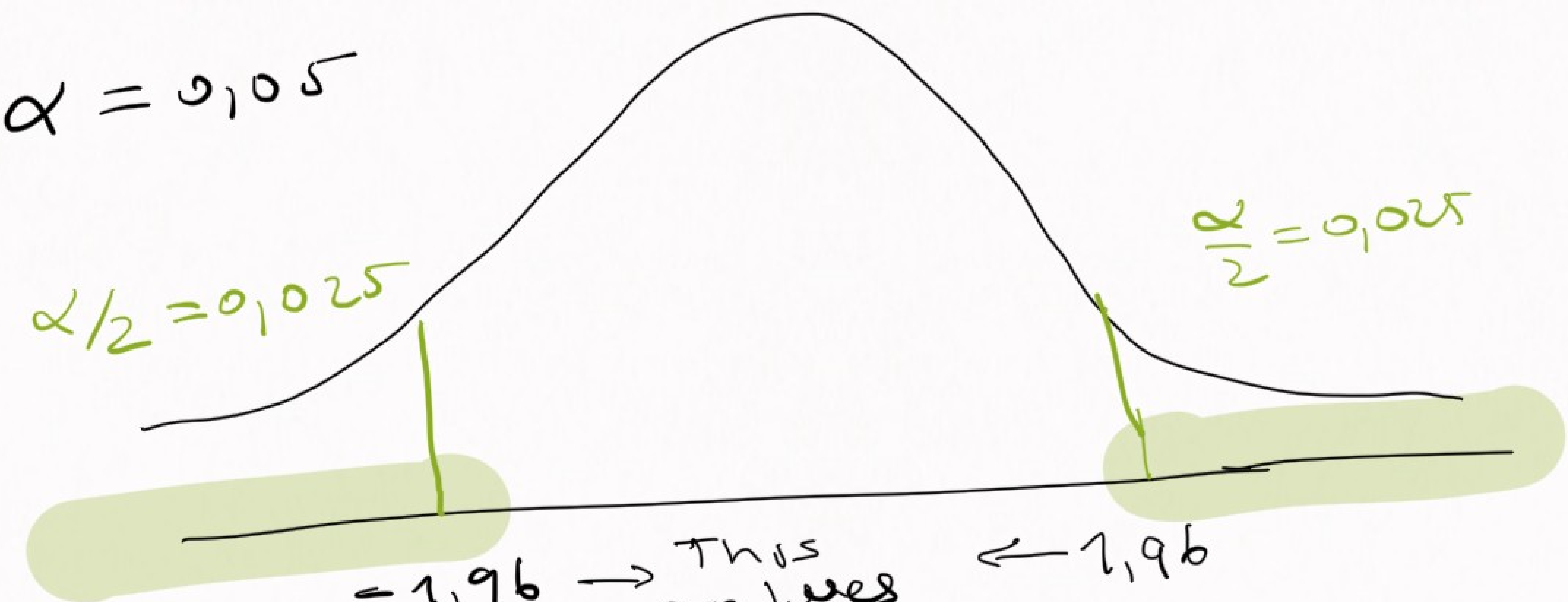
- We're scaling or standarizing the sample mean.
- If the sample mean is close to the population one $\Rightarrow Z \rightarrow 0$ (close to zero)

**How big should be Z to reject $H_0$?**

distribution of Z
(standard normal distribution)



- When we compute $Z$, we get a value, If this value falls in the blue area we accept $H_0$. Otherwise, we reject.
- The rejection region depends on $\alpha$. (level of significance).




$-1{,}96 \rightarrow$ This values come from ~~the~~ Z-table. $\leftarrow 1{,}96$

## How does hyp. testing work?

- Calculate a statistic (eg. $\bar{x}$ mean) from your sample
- Scale it. (eg. $Z_1 = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$)
- Check if $Z_1$ is in the rejection region.

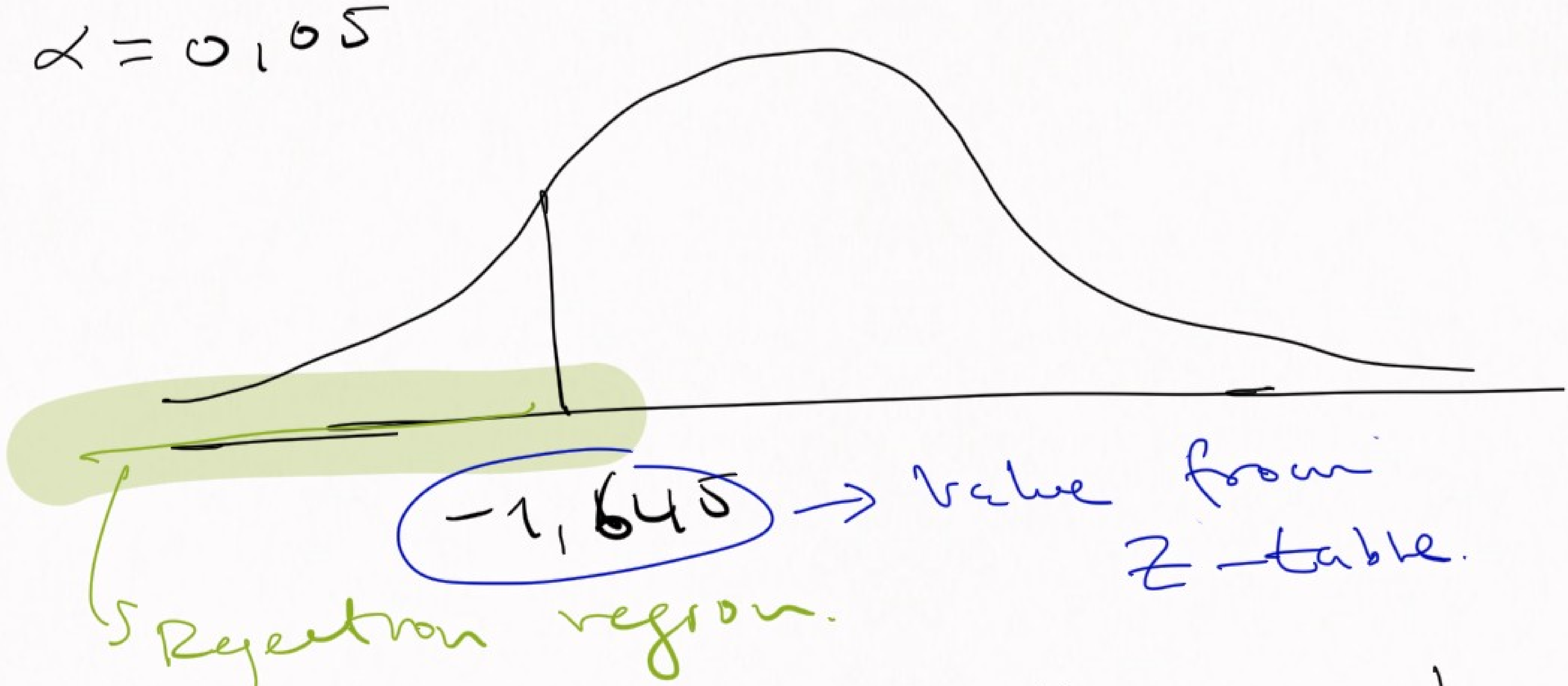
## What about. 1-sided Tests?

# ONE SIDED TESTS

$H_0 := \mu_0 \geq 125{,}000$ USD.

$H_1 := \mu_0 < 125{,}000$ USD

$\alpha = 0{,}05$




If after computing $Z$ we get that $Z < -1{,}645$, we would reject the null hypothesis

# Errors in Hyp. testing

- **Type I** : (False positive).
  Rejecting $H_0$ when its true.

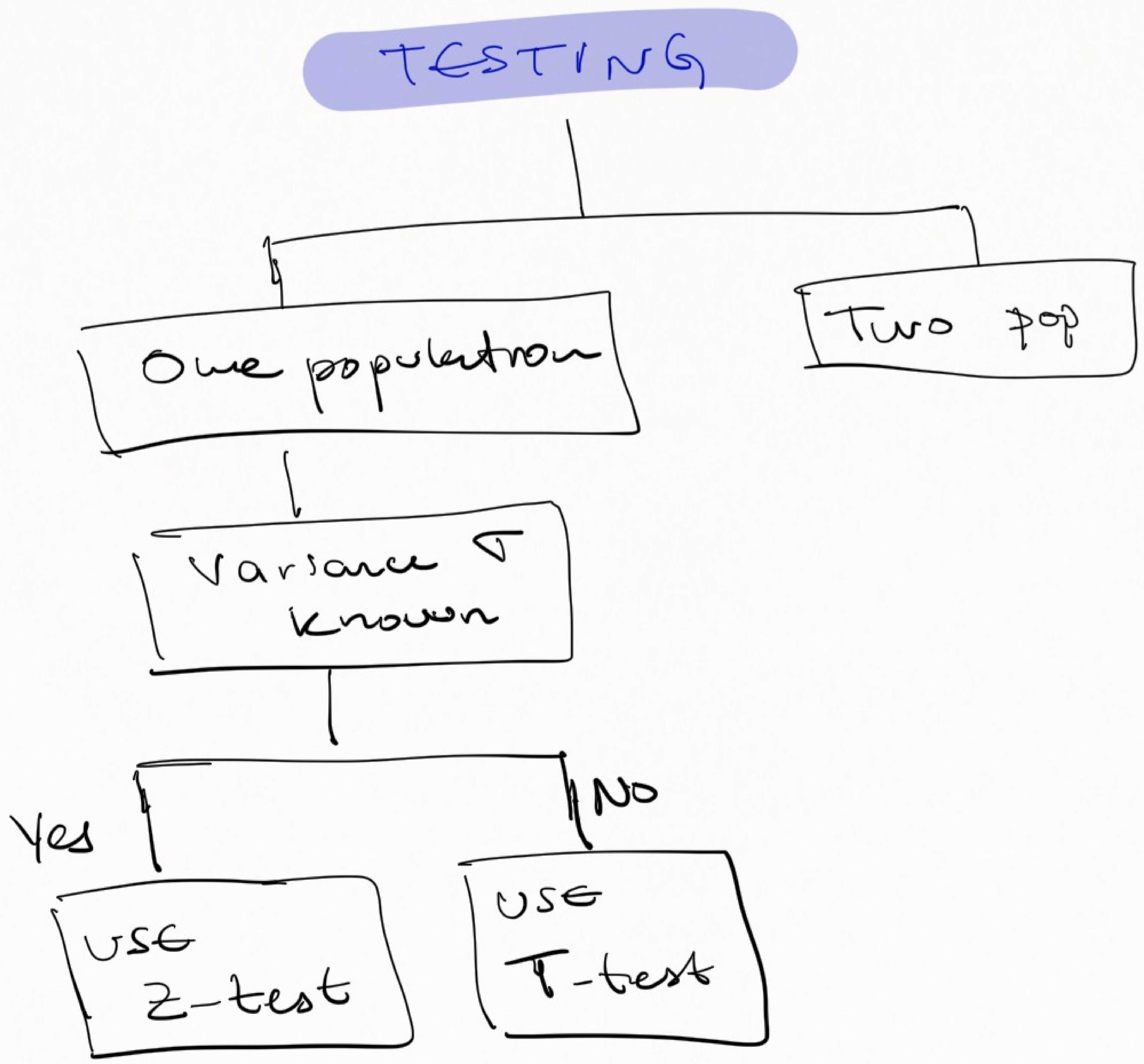
  probability : $\alpha$.

- **Type II** : (False negative).
  Accepting $H_0$ when False

  prob: $\beta$

- Rejecting a false $H_0$.
  prob : $1-\beta$ !!! (AKA. Power of the test)

  - test goals reject all FALSE $H_0$




<u>Example</u> "Data science Salary" (AGAIN)

<u>Sample mean</u> ($\bar{x}$) \$ 100,200 USD

<u>pop. std.</u> ($\sigma^2$) \$ 15,000

<u>standard error</u> $\sigma$ \$ 2,739

<u>Sample size</u> 30

Glassdoor \$113,000

$H_0 : \mu_0 = \$113{,}000$

$H_1 : \mu_0 = \$113{,}000$

Test for the mean, pop. variance known

$$Z_1 = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}} \sim N(0,1)$$

$$= \frac{100{,}200 - 113{,}000}{2{,}739} = -4{,}67.$$

- Now we compare $|-4{,}67|$ with $z_{\alpha/2}$.
- Some z-tables don't include negative values.
- As the standard normal distribution is symmetrical around 0, the two statements are equivalent:

$-4{,}67 <$ A negative z.

is equivalent to:

$4{,}67 >$ A positive z.

## Decision Rule

Reject if: $|Z| > z$
absolute value of Z-score > positive crucial value z.

$4{,}67 > 1{,}96.$

$\Rightarrow$ we reject $H_0$.

w/ $\alpha = 0{,}05$ and assuming $H_0$ true

First test done!!! :)

So far we know ... (Do we?)

- How to test hypothesis
- How to reject them
- ... at various levels of significance

- But we don't know
  - A level of significance after which we can no longer do it.

## P-VALUE

the smallest level of significance at which we can still reject the null hypothesis, given the observed sample statistic.

Example "Glassdoor salaries" AGAIN (x2)

$s$ = std. error = 2739

$\sigma$ = pop. std = 15000

The salaries are normally distributed

$N \sim (\mu, \sigma^2)$, $\boxed{n = 30}$ sample size

$Z = -4{,}67$

we reject at $\alpha = 0{,}05$ and $\alpha = 0{,}01$

- p-value $= 1 - (\#$ from the table$)$
$= 0{,}001$

- In the table (Z-table) we will not find the value we're looking for, so we take the closest one and from the table is -

$\boxed{3{,}0 + 0{,}09} \Rightarrow \boxed{0{,}9990}$ crucial/critical value.

<u>RULE</u>: Reject Ho if.
p-value $< \alpha$

$Z = -4{,}67$;
p-value for $Z = 0{,}0001$

$$\left.\begin{array}{l} \text{Test at } 90\%: 0{,}0001 < 0.1 \\ \text{Test at } 95\%: 0{,}0001 < 0{,}05 \\ \text{Test at } 99\%: 0{,}0001 < 0{,}01 \end{array}\right\}$$

<u>General conclusion</u>

Reject the null hypothesis

08.12.2020