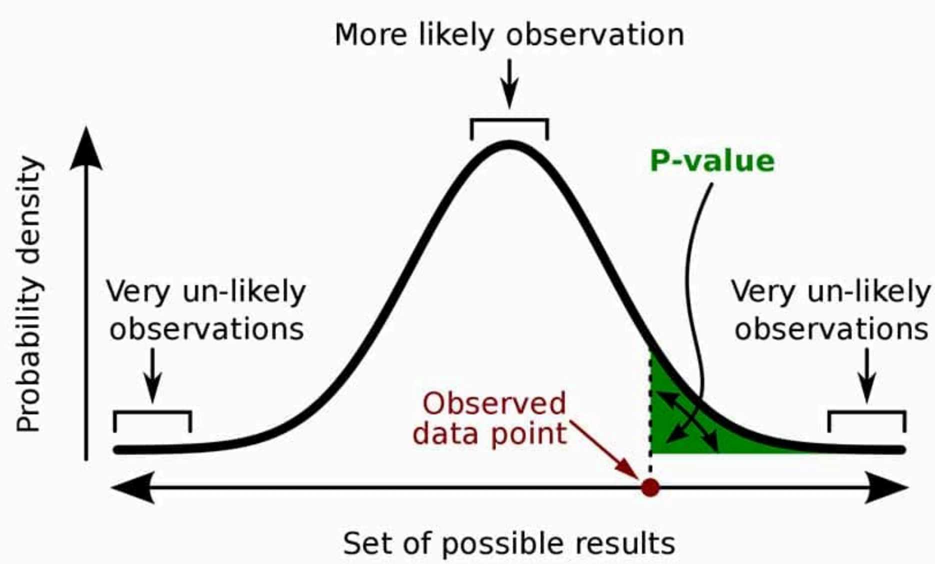
Hypothesis Testing Part 2 | p-values | t-tests

P-value (VVIMP for Interviews)

- **p-value** is the probability of obtaining results as extreme as, or more extreme than, what we observed in our sample data, **assuming the null hypothesis (H_o) is true**.

It quantifies the evidence against the null hypothesis:

- A **small p-value** (typically less than 0.05) indicates that the observed data is very unlikely under H_o, suggesting that **H_o should be rejected**.
- A **large p-value** indicates that the observed data is consistent with H_o, so we d**o not have sufficient evidence to reject it**.



A **p-value** (shaded green area) is the probability of an observed (or more extreme) result assuming that the null hypothesis is true.

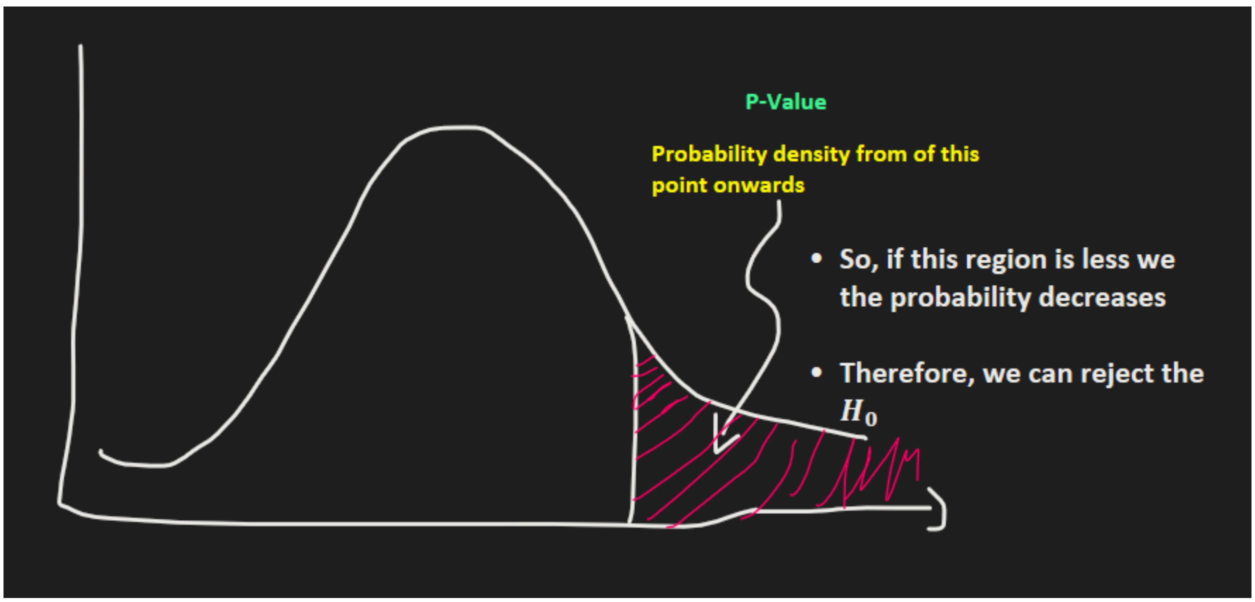
Decision Rule:

- If **p-value < α** (commonly **0.05**), we reject H_o.
- If **p-value ≥ α**, we fail to reject H_o.

Example

Imagine you have a coin that is supposed to be fair (H_o: the coin lands on heads 50% of the time). You flip the coin 10 times and get 9 heads. Now, the p-value tells you:

- "If the coin were really fair, what is the chance of getting 9 or more heads out of 10 flips?"
- If that chance (p-value) is very small, you have evidence against the coin being fair.

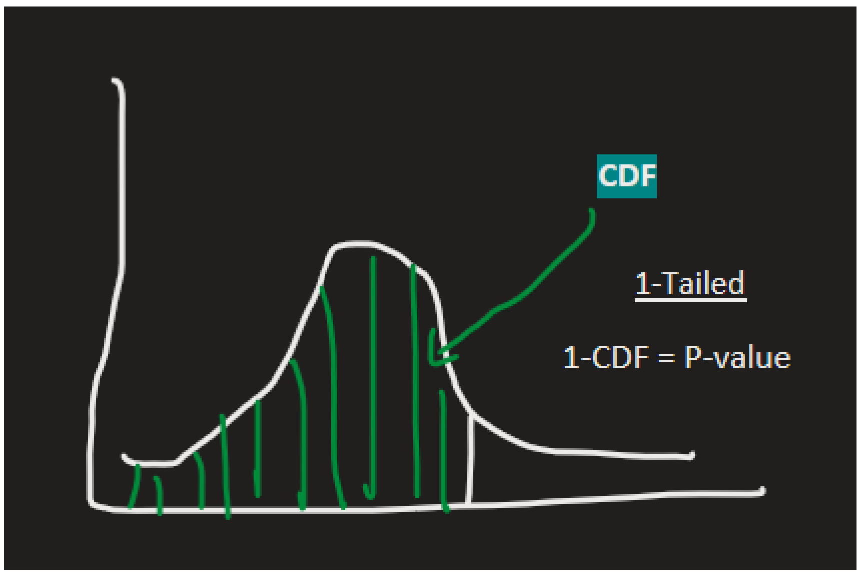


How to calculate P-Value?

1. First we calculate the **t-statistic with the formula**

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

2. Then we use the t-distribution (with degrees of freedom n−1) to find the p-value
    - `stats.t.cdf(abs(t_statistic), df=n-1)` gives the cumulative probability up to our computed t-value.

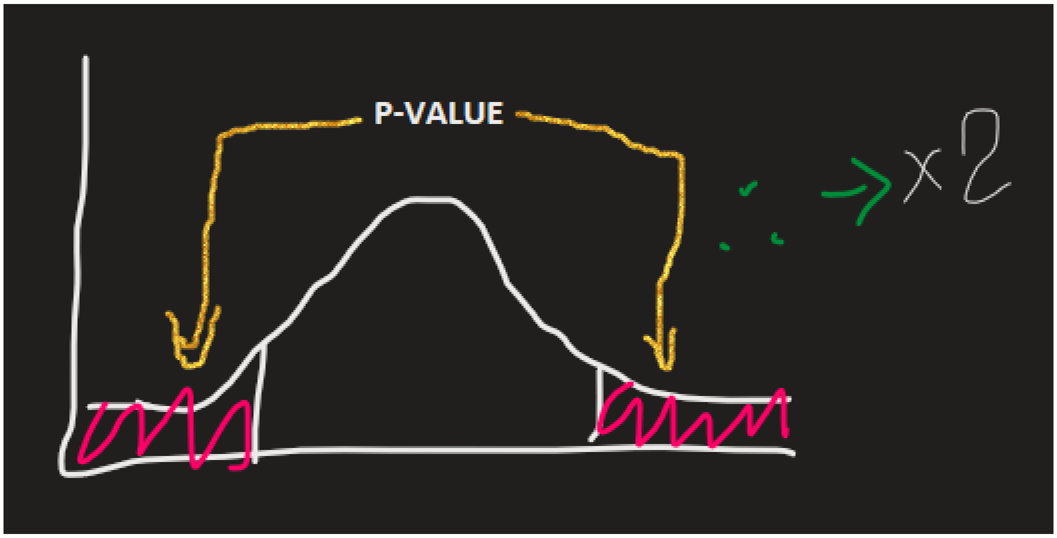


- p-value = `1 -` `stats.t.cdf(abs(t_statistic), df=n-1)`
    - Subtract from 1

3. For **2-tailed,** multiply by 2

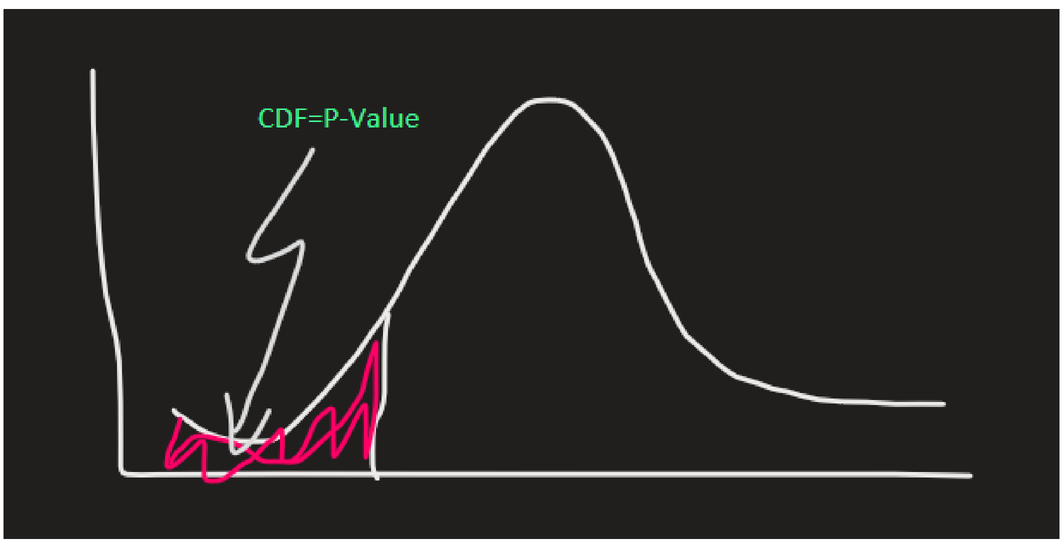
   `2 *` `(1 - stats.t.cdf(abs(t_statistic), df=n-1))`

$$p - value = 2 \times Tail\ Area$$



For left-tailed test:

- $p-value = CDF(observedvalue)$



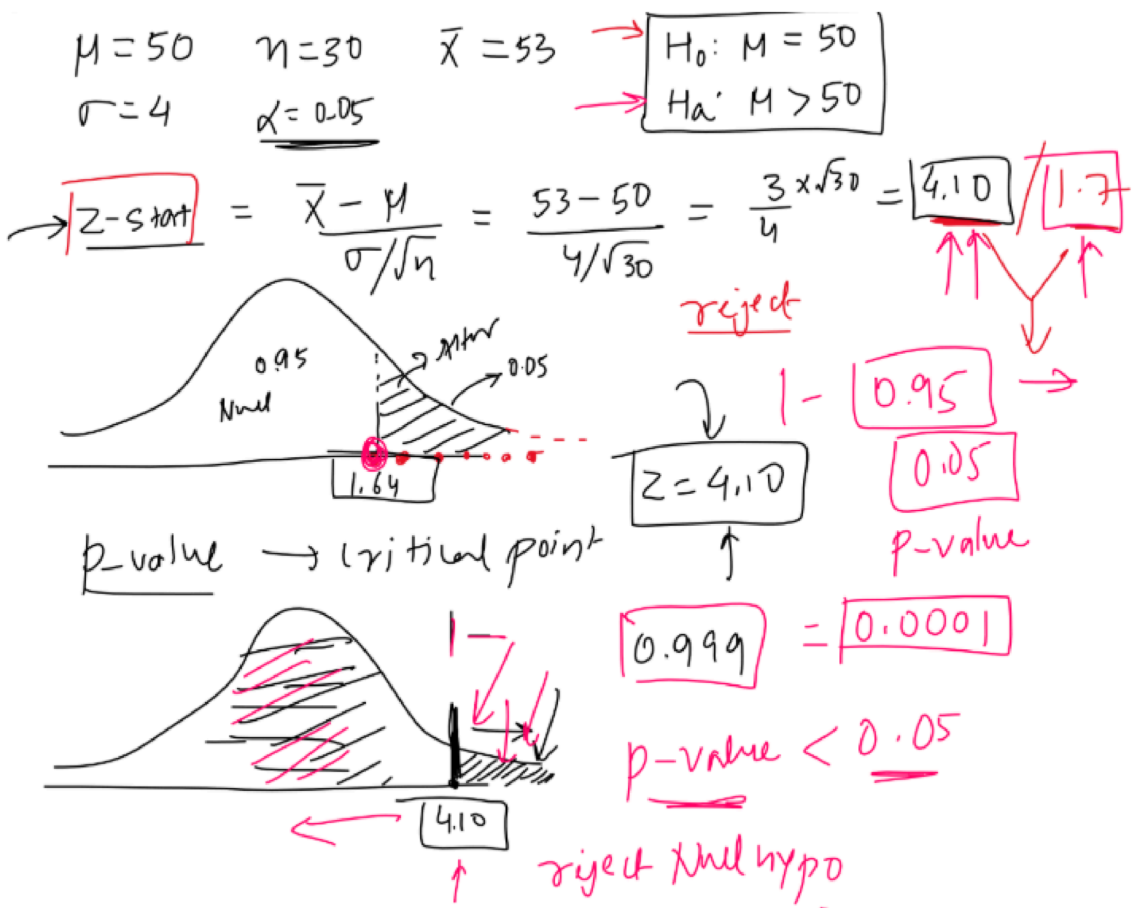
Without significance value

- Very small p-values (e.g., p < 0.01) indicate strong evidence against the null hypothesis
- Small p-values (e.g., 0.01 ≤ p < 0.05) indicate moderate evidence
- Large p-values (e.g., 0.05 ≤ p < 0.1) indicate weak evidence

- Very large p-values (e.g., p ≥ 0.1) indicate weak or no evidence

P-value in context of Z-test

Q. Suppose a company is evaluating the impact of a new training program on the productivity of its employees. The company has data on the average productivity of its employees before implementing the training program. The average productivity was 50 units per day. After implementing the training program, the company measures the productivity of a random sample of 30 employees. The sample has an average productivity of 53 units per day and the pop std is 4. The company wants to know if the new training program has significantly increased productivity.



- Here, you don't calculate the ~~**critical value.**~~
- You calculate **z/t statistic** as usual & after that:
    - You calculate **p-value**→ The **area to the right of z/t stat**
- Here, z=4.10
- The area to the right of 4.10
    - Find this from z-table = 0.999 (This is area to the right)
    - To get p-value, we have to subtract this value from 1
- p value = 1-0.999 = **0.0001**

Here,

p value<α value → Reject the H_0

Q. 2-Tailed

Suppose a snack food company claims that their Lays wafer packets contain an average weight of 50 grams per packet. To verify this claim, a consumer watchdog organization decides to test a random sample of Lays wafer packets. The organization wants to determine whether the actual average weight differs significantly from the claimed 50 grams. The organization collects a random sample of 40 Lays wafer packets and measures their weights. They find that the sample has an average weight of 49 grams, with a pop standard deviation of 5 grams.

$\mu = 50$ $n = 40$ $\bar{x} = 49$ $H_0: \boxed{\mu = 50}$

$\sigma = 5$ $\alpha = 0.05$ $H_a: \mu \neq 50$

p-value 2tailed $\boxed{0.206} > 0.05$

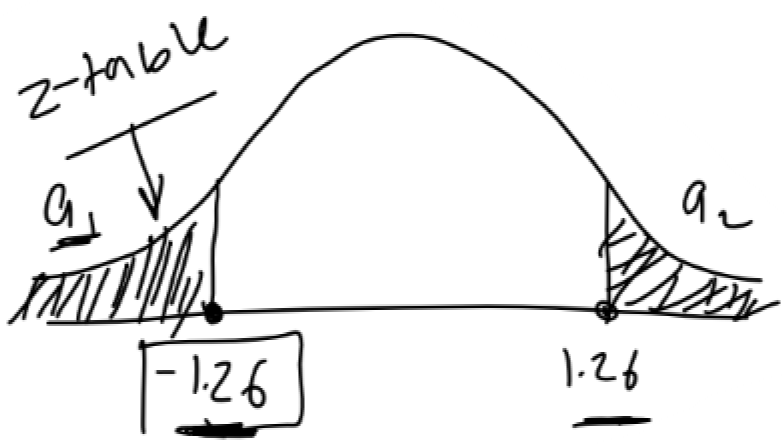
$$Z = \frac{49 - 50}{5/\sqrt{40}} = \frac{-\sqrt{40}}{5} = \boxed{-1.26}$$

$\boxed{\text{p-val}}$ $\boxed{0.206}$

$\text{p-value} = a_1 + a_2$

Z= -1.26

We have to take the are area left to -1.26 and right to +1.26 as this is 2-tailed test



p-value will be $2 * area$

STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.

Z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
-3.9	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
-3.8	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
-3.7	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
-3.6	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
-3.5	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
-3.4	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
-3.3	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
-3.2	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
-3.1	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
-3.0	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
-2.9	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
-2.8	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
-2.7	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
-2.6	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
-2.5	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
-2.4	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
-2.3	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
-2.2	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
-2.1	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
-2.0	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
-1.9	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
-1.8	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
-1.7	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
-1.6	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
-1.5	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
-1.4	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
-1.3	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
-1.2	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
-1.1	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
-1.0	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
-0.9	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109

Total area = 0.103*2 = **0.206**

P-Value = 0.206

P-Value>α (0.05) → We cannot Reject the H_0

T-test

A **t-test** is a statistical test that compares the means (averages) of two groups to see if they are significantly different from each other.

- You use for smaller samples
- When you don't have population mean (typically n < 30).

Types of t-tests

1. **One-Sample t-test:**
    - Compares the **mean of a single sample to a known value** (usually a population mean).
    - *Example:* Testing if the average weight of apples is 150 grams.
2. **Independent Two-Sample t-test:**
    - Compares the **means of two independent groups**.
    - *Example:* Comparing test scores between two different classrooms.
3. **Paired t-test (Dependent t-test):**
    - Compares the means of the **same group** at two different times or under two different conditions.
    - *Example:* Measuring blood pressure before and after a treatment on the same group of patients.

Hypotheses Setup

- **Null Hypothesis (H_o):** There's no difference in means.
    - One-sample: $H_0 : \mu = \mu_0$
    - Two-sample: $H_0 : \mu_1 = \mu_2$
- **Alternative Hypothesis (H_1):** There is a difference.
    - One-sample: $H_1 : \mu \neq \mu_0$
    - Two-sample: $H1 : \mu_1 \neq \mu_2$

Test Statistic Calculation

For a one-sample t-test:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

- $\bar{x}$: Sample mean.
- μ_0: Population mean (hypothesized value).
- s: Sample standard deviation.
- n: Sample size.

For a two-sample t-test (assuming equal variances):

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

- $\bar{x}_1, \bar{x}_2$: Means of the two samples.
- s_p^2: Pooled variance.
- n_1, n_2: Sample sizes.

Decision Rule

If p-value < α, reject H_o.
If
p-value ≥ α, do not reject H_o.

```
import numpy as np
from scipy import stats

# Sample data
test_scores = np.array([72, 78, 74, 80, 76, 79, 71, 77, 73, 75])

# Sample statistics
sample_mean = np.mean(test_scores)
sample_std = np.std(test_scores, ddof=1)  # ddof=1 for sample standard devia
tion
n = len(test_scores)
mu0 = 75  # Hypothesized population mean

# t-test calculation
t_statistic, p_value = stats.ttest_1samp(test_scores, mu0)

# Output results
print(f"Sample Mean: {sample_mean:.2f}")
print(f"t-Statistic: {t_statistic:.2f}")
print(f"P-value: {p_value:.4f}")

# Decision
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis: The average score is significantly different
from 75.")
else:
    print("Fail to reject the null hypothesis: There is no significant difference fro
m 75.")
```

Output:

Sample Mean: 75.50
t-Statistic: 0.52
P-value: 0.6141
Fail to reject the null hypothesis: There is no significant difference from 75.

Calculate are to the left (CDF) of T-Distribution

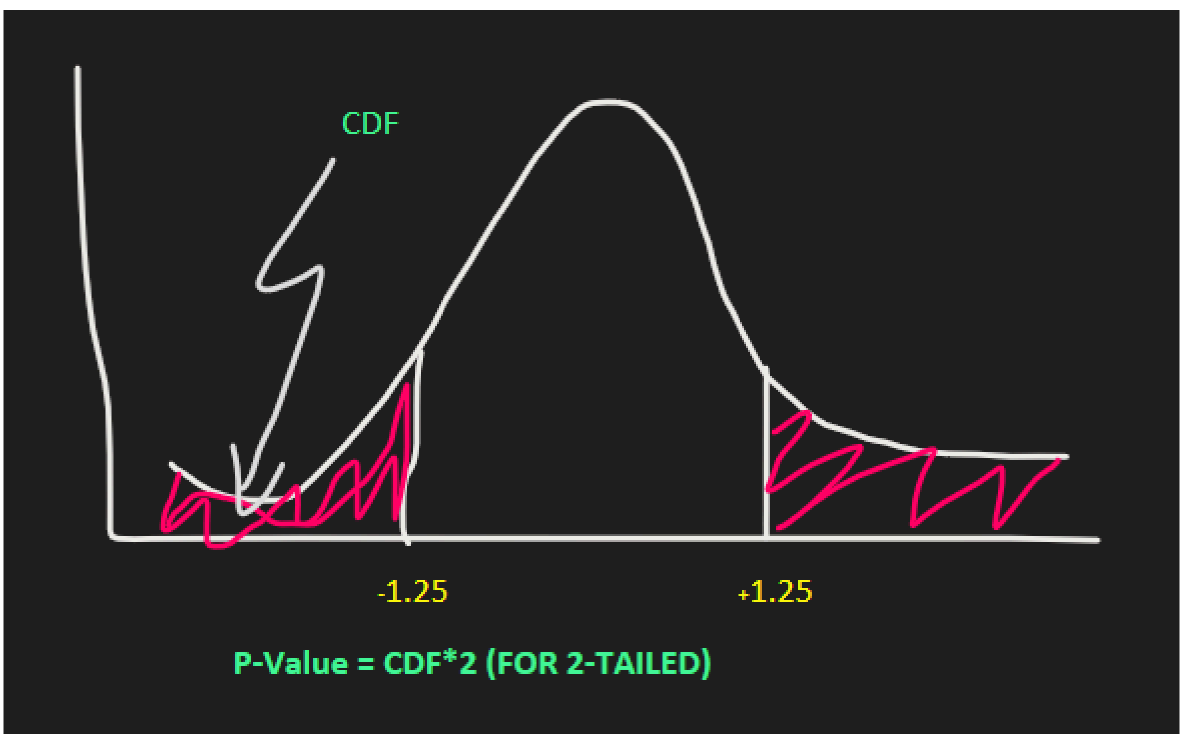
```
from scipy.stats import t

# Set the t-value and degrees of freedom
t_value = -1.25
df = 58  # Replace this with your specific degrees of freedom

# Calculate the CDF value
cdf_value = t.cdf(t_value, df)
print(cdf_value*2)

Output: 0.21631974483731498
```

we did `cdf_value*2` because it was 2-tailed test



T-Test using Python

```
from scipy.stats

stats.ttest_1samp(sample_data, population_mean)
```

Returns:

1. **t-statistic**
2. **p-value**

Example:
If you have a sample of heights and want to test if the average height differs from 170 cm (the population mean), you can do:

```
from scipy import stats

sample = [172, 173, 170, 171, 175, 169, 168, 171, 174, 172, 170, 173, 169, 168, 175]
pop_mean = 170

t_stat, p_value = stats.ttest_1samp(sample, pop_mean)
print("t-statistic:", t_stat)
print("p-value:", p_value)
```

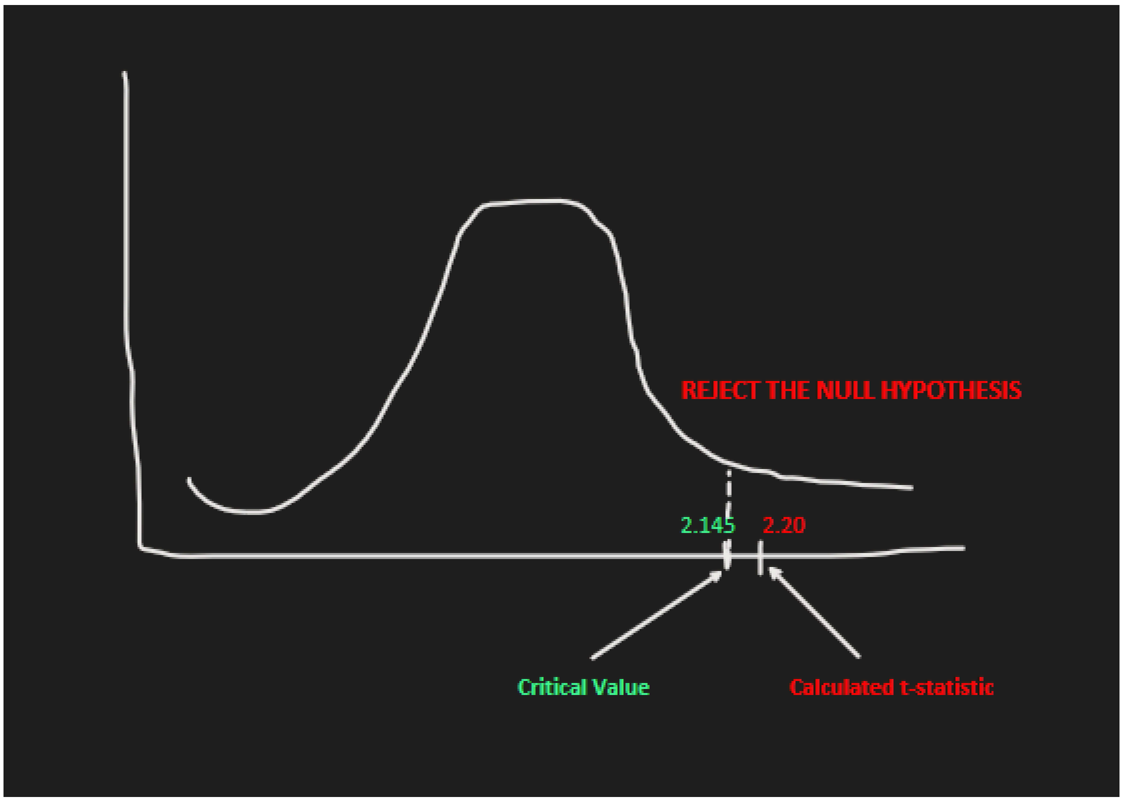
Output:

t-statistic: 2.1971768720102216
p-value: 0.045339544447787296

- the **p-value** is **0.0453**, which is the probability of observing a t-statistic as extreme as 2.20 in either direction (**both tails combined**).

- **Calculated t-statistic**: 2.20
- **Critical t-values**: ±2.145

Since **2.20 > 2.145**, the **calculated t-statistic** lies **outside the range** of the critical values, so you **reject the null hypothesis**.



Independent 2 sample t-test

- The independent two-sample t-test (also called an unpaired t-test) is used to compare the means of **two independent groups** to determine if there is a significant difference between them.

Formula:

The t-statistic is calculated as:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$$

Where:

- $\bar{X}_1, \bar{X}_2$ = Sample means of groups 1 and 2
- n_1, n_2 = Sample sizes of groups 1 and 2
- s_p^2 = Pooled variance, calculated as:

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

- s_1^2, s_2^2 = Sample variances of groups 1 and 2

The test assumes **equal variances** by default but can be adjusted for unequal variances using Welch's t-test.

Assumptions:

1. **Independence:** The two samples are independent of each other.
2. **Normality:** The data in each group should be approximately normally distributed (especially for small samples).
3. **Equal or Unequal Variances:**
    - If variances are equal → Use the standard t-test formula.
    - If variances are unequal → Use **Welch's t-test** (handled automatically by `scipy.stats.ttest_ind()` using `equal_var=False`).

```
import numpy as np
from scipy import stats

# Generate random samples for two independent groups
np.random.seed(42)
group1 = np.random.normal(loc=170, scale=5, size=30)  # Mean=170, Std=5, n
=30
group2 = np.random.normal(loc=175, scale=6, size=35)  # Mean=175, Std=6,
n=35

# Perform an independent t-test
t_stat, p_value = stats.ttest_ind(group1, group2, equal_var=True)  # Assuming e
qual variances

# Output results
print("t-statistic:", t_stat)
print("p-value:", p_value)

# Check significance at α = 0.05
alpha = 0.05
if p_value < alpha:
    print("Reject the null hypothesis: The means are significantly different.")
else:
    print("Fail to reject the null hypothesis: No significant difference in means.")
```

Output:

t-statistic: -3.943177666481369
p-value: 0.00020453756573145986
Reject the null hypothesis: The means are significantly different.

Q. Suppose a website owner claims that there is no difference in the average time spent on their website between desktop and mobile users.

To test this claim, we collect data from 30 desktop users and 30 mobile users regarding the time spent on the website in minutes. The sample statistics are as

follows:

```
desktop_users = [12, 15, 18, 16, 20, 17, 14, 22, 19, 21, 23, 18, 25, 17, 16, 24, 20, 19, 22, 18, 15, 14, 23, 16, 12, 21, 19, 17, 20, 14]

mobile_users = [10, 12, 14, 13, 16, 15, 11, 17, 14, 16, 18, 14, 20, 15, 14, 19, 16, 15, 17, 14, 12, 11, 18, 15, 10, 16, 15, 13, 16, 11]
```

Desktop users:

- Sample size (n1): 30
- Sample mean (mean1): 18.5 minutes
- Sample standard deviation (std_dev1): 3.5 minutes

Mobile users:

- Sample size (n2): 30
- Sample mean (mean2): 14.3 minutes
- Sample standard deviation (std_dev2): 2.7 minutes

We will use a significance level (α) of 0.05 for the hypothesis test.

Shapiro-Wilk test for checking Normality:

```
desktop_users = [12, 15, 18, 16, 20, 17, 14, 22, 19, 21, 23, 18, 25, 17, 16, 24, 20, 19, 22, 18, 15, 14, 23, 16, 12, 21, 19, 17, 20, 14]
mobile_users = [10, 12, 14, 13, 16, 15, 11, 17, 14, 16, 18, 14, 20, 15, 14, 19, 16, 15, 17, 14, 12, 11, 18, 15, 10, 16, 15, 13, 16, 11]

# Perform the Shapiro-Wilk test for both desktop and mobile users
shapiro_desktop = shapiro(desktop_users)
shapiro_mobile = shapiro(mobile_users)
```

```
print("Shapiro-Wilk test for desktop users:", shapiro_desktop)
print("Shapiro-Wilk test for mobile users:", shapiro_mobile)
```

Output:

Shapiro-Wilk test for desktop users: ShapiroResult(statistic=0.9783115512411942, pvalue=**0.7791003299808725**)

Shapiro-Wilk test for mobile users: ShapiroResult(statistic=0.9714355768676655, pvalue=
0.5791606602037616)

- Both are more than 0.05. Therefore, are normal.

We will use Levene's test to check the variance:

```
# Perform Levene's test

from scipy.stats import levene

levene_test = levene(desktop_users, mobile_users)
print(levene_test)

Output:
LeveneResult(statistic=2.94395488191752, pvalue=0.09153720526741756)
```

- **p-value ≥ 0.05**→ We fail to reject the H_0 → **Variances are equal** → Use standard t-test.

Now calculate the P value with independent 2 sample test:

`stats.` `ttest_ind` **()**

```
t_stat, p_value = stats.ttest_ind(desktop_users, mobile_users)
print("t-statistic:", t_stat)
print  ("p-value:", p_value)

Output:
t-statistic: 4.625335930681123
p-value: 2.1422811334975257e-05
```

- p-value<α(0.05) → Reject the H_0
  - P-value is ***2.14 x e-05***
- There is ***difference in the average time spent on website between desktop and mobile users.***

Paired 2 sample t-test

A **paired t-test** (dependent t-test) is used to compare **two related (paired) samples** to determine if their means are significantly different.

🔷 **Key Use Case:** When the same group is measured **before and after** an intervention.

When to Use a Paired t-Test?

Use a **paired t-test** when:

- The **same subjects** are measured twice (e.g., pre-test vs. post-test).
- There is a **natural pairing** (e.g., twin studies, left hand vs. right hand).
- The data are **normally distributed** (for small samples, say n<30).

  n<30n < 30

Hypotheses

- **Null Hypothesis (H_0):** There is **no significant difference** in the mean values before and after.H0:μd=0

  $$H_0 : \mu_d = 0$$

  (where μ_d is the mean of the differences)

- **Alternative Hypothesis (H_1):** There **is a significant difference** in the mean values.

  $$H_1 : \mu_d \neq 0$$

Formula for Paired t-Test

$$t = \frac{\bar{d}}{s_d / \sqrt{n}}$$

Where:

- $\bar{d}$ = Mean of the differences
- s_d = Standard deviation of the differences
- n = Number of paired samples

The **p-value** is then obtained using the t-distribution.

Formula for s_d

$$s_d = \sqrt{\frac{\sum(d_i - \bar{d})^2}{n-1}}$$

Where:

- d_i = Individual differences between paired observations
- $\bar{d}$ = Mean of the differences
- n = Number of paired samples
- s_d = Standard deviation of the differences

Degrees of Freedom

$n - 1$

Python Implementation

```
import numpy as np
from scipy import stats

# Sample data: Before and After scores of the same individuals
before = np.array([72, 75, 78, 79, 80, 82, 85, 88, 90, 91])
after  = np.array([74, 76, 76, 80, 82, 81, 86, 89, 92, 93])

# Perform a paired t-test
t_stat, p_value = stats.ttest_rel(before, after)

# Output results
print(f"t-statistic: {t_stat:.4f}")
```

```
print(f"p-value: {p_value:.4f}")

# Decision at α = 0.05
alpha = 0.05
if p_value < alpha:
    print("Reject H₀: There is a significant difference between before and afte
r.")
else:
    print("Fail to reject H₀: No significant difference.")
```

Output:

t-statistic: -2.0769
p-value: 0.0676
Fail to reject H_0: No significant difference.

Interpretation

- **p-value < 0.05** → Reject H0 → Significant difference exists.
- **p-value ≥ 0.05** → Fail to reject H0 → No significant difference.