Advanced Statistics

DigitalSkola
Uncover The World Of Digital Skills With Us

Quote of the day

A Data Scientist is one who knows more statistics than a programmer and more programming than a statistician

- Josh Wills

DigitalSkola
Uncover The World Of Digital Skills With Us

Hello!
I am Agil Haykal

I am a Data expert with extensive experience in multiple industries such as marketplace, insurance, banking, general taxation, consulting, and training.

In total, I trained more than 300 data scientists, engineers, and analysts.

DigitalSkola
Uncover The World Of Digital Skills With Us

Table of Content

What will We Learn Today?

1. Sampling Methods
2. Hypothesis testing
3. AB Testing

DigitalSkola
Uncover The World Of Digital Skills With Us

Sampling

Why we need Sampling?

1. Analyze the population would be **costly**.
2. Take population data **consume a very long time**.
3. **Computational cost** also proportionate the data size.
4. Sometimes we are **limited to get** the population data.

DigitalSkola
Uncover The World Of Digital Skills With Us

Types of Sampling Method

Probability Sampling

Non-Probability Sampling

What is Probability Sampling?

Probability sampling is a sampling technique that let **every sample has equal chance.** Consequently, we can cherry pick sample that we want to use.

What is Non-Probability Sampling?

Probability sampling is a sampling technique that we can **set certain rule** to get the sample. To put it simply, we can get the sample **by accident**.

Type of Sampling Errors

Sampling Error

Mistake caused by observing wrong population or define wrong population. We can reduce this error by clearly define the population and design the sampling method properly.

Non-Sampling Error

Error caused by external factors that we cannot control.

Sampling Size

Slovin :

$$n = \frac{N}{1 + Ne^2}$$

N : Number of Population
e : Margin of Error

Use Case of Sampling

Indonesian Survey organization is going to survey Pemilu 2019.
Given the Indonesian population is 200.000.000. How many sample is needed when we allow margin of error 1%?

Simple random sampling

Sample 400 observation

```
df.sample(n=400)
```

	user_id	timestamp	group	landing_page	converted
84561	889913	2017-01-10 03:29:44.877020	control	old_page	1
158240	795331	2017-01-05 02:28:31.324120	control	old_page	0
16396	931481	2017-01-14 10:27:44.036521	treatment	new_page	0
214735	678545	2017-01-10 20:09:08.870690	control	old_page	1
270680	938785	2017-01-05 16:55:07.131904	treatment	new_page	1
...	...	...	...	...	...
189394	634341	2017-01-20 12:30:19.787521	control	old_page	0
49744	873063	2017-01-07 08:50:00.929656	treatment	new_page	1
158707	783264	2017-01-06 06:11:55.341772	treatment	new_page	0
110721	870153	2017-01-11 14:05:31.435297	treatment	new_page	0
218199	703692	2017-01-24 05:53:55.056226	control	old_page	0

400 rows × 5 columns

Simple random sampling

Sample 30% from dataset

```
df.sample(frac=0.3)
```

	user_id	timestamp	group	landing_page	converted
186583	893743	2017-01-11 04:39:37.172078	control	old_page	0
256031	853511	2017-01-03 05:00:45.694946	control	old_page	0
283006	814654	2017-01-05 21:04:18.990880	treatment	new_page	0
254110	751639	2017-01-08 14:59:31.537678	control	old_page	0
181889	665122	2017-01-06 17:44:30.179297	control	old_page	0
...	...	...	...	...	...
52803	670220	2017-01-06 11:44:33.323235	control	old_page	0
26239	888236	2017-01-23 05:55:08.816955	treatment	new_page	0
47497	845510	2017-01-03 16:58:13.046340	control	old_page	0
77750	653130	2017-01-22 18:56:55.587907	treatment	new_page	0
119070	785120	2017-01-14 05:27:37.836025	control	old_page	0

88343 rows × 5 columns

Stratified random sampling

Sampling for Each Group

```
df.groupby(['group'],as_index=True).apply(lambda x: x.sample(n=200,random_state=123))
```

		user_id	timestamp	group	landing_page	converted
group						
control	95574	704344	2017-01-08 06:33:15.620318	control	old_page	0
	282637	903218	2017-01-07 16:40:31.904242	control	old_page	0
	201262	724634	2017-01-05 18:38:31.257679	control	old_page	0
	93315	750623	2017-01-21 19:20:32.814948	control	old_page	0
	16163	651056	2017-01-04 03:17:39.846424	control	old_page	0
...	...	...	...	...	...	...
treatment	16034	665227	2017-01-18 06:10:37.832101	treatment	new_page	1
	241972	818984	2017-01-23 01:45:24.506789	treatment	new_page	0
	135298	843757	2017-01-04 03:10:19.433517	treatment	new_page	0
	200501	659763	2017-01-24 13:21:56.026713	treatment	new_page	0
	158648	788418	2017-01-14 05:09:32.246838	treatment	new_page	0

400 rows × 5 columns

DigitalSkola
Uncover The World Of Digital Skills With Us

Hypothesis Testing

Previous Topic

After we know descriptive statistics form our data (mean, median, st dev, mode, etc), so what next?

DigitalSkola
Uncover The World Of Digital Skills With Us

Get Conclusions!

DigitalSkola
Uncover The World Of Digital Skills With Us

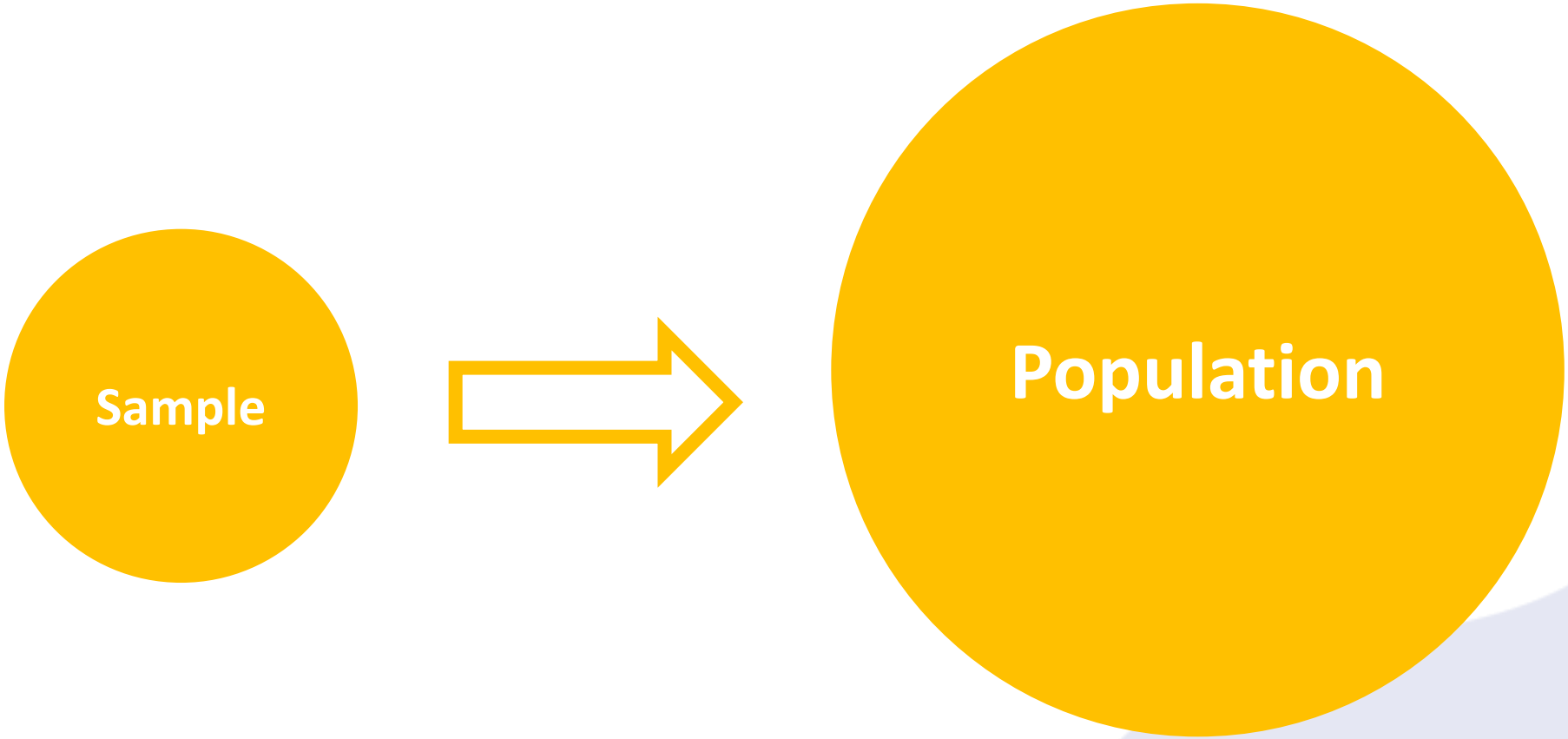
How to get conclusions?

DigitalSkola
Uncover The World Of Digital Skills With Us

Inference Statistics

DigitalSkola
Uncover The World Of Digital Skills With Us

Inference Statistics



Diet Plan

We want to prove which diet plan works best

Non-supplement Diet

Duration: **3 months**

Avg weight drop: **8.1 kg**

St Dev: **2.3 kg**

Supplement Diet

Duration: **3 months**

Avg weight drop: **8.5 kg**

St Dev: **1.5 kg**

Which one is better?

Objective and Principle of Hypothesis Testing

Objective of hypothesis testing is **to prove assumptions** by using existing sample data.

Its principle is **presumption of innocence** (praduga tak bersalah).

Illustration

Is he the perpetrator?

Hypothesis

Hypothesis 0: He is not the thief

Hypothesis 1: He is the thief

Evidences:

Data:

1. CCTV shows **he is at the location**.
2. **There is a witness** who claimed He is there.
3. There is **his fingerprint** nearby.

The Verdict

Based on those evidences, He is **strongly suspected** as a Thief.

Because risk of making him an innocence is very small.

Then we can conclude he is the thief. (accepting Hypothesis 1).

Hypothesis

Hypothesis 0: He is not the thief

Hypothesis 1: He is the thief

Hypothesis testing in Statistics

In Statistics, Hypothesis 0 that we used previously usually called as Null Hypothesis.

Null Hypothesis (H_0)

Is a hypothesis that contradicts the assumption we are going to test. And its characteristics is general.

Hypothesis testing in Statistics

In Statistics, Hypothesis 1 that we used previously usually called as Alternative Hypothesis.

Alternative Hypothesis (H_a /H_1)

Is a hypothesis that we going to test. And its characteristics is specific.

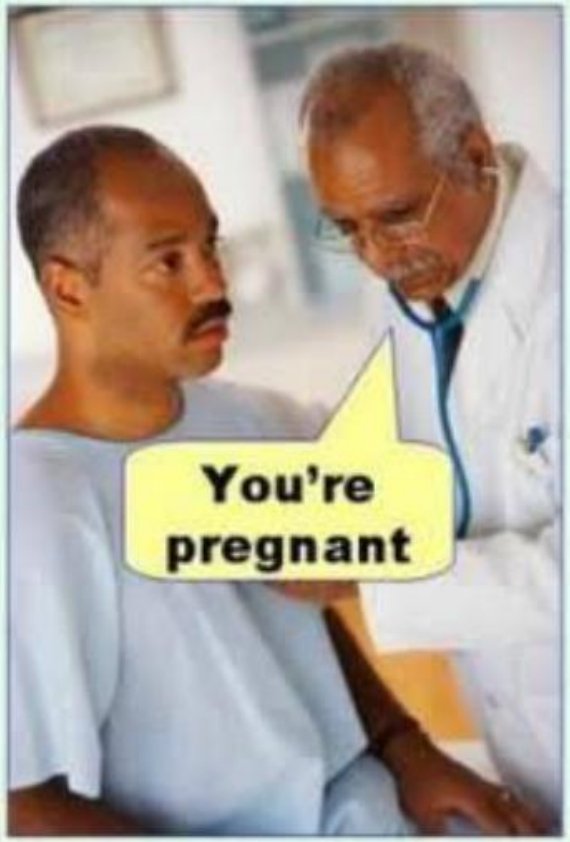
Possible error of Hypothesis Testing

	Not a Thief	A Thief
Not Guilty	Proven Not Guilty Not a thief	Proven not Guilty Really a thief (β)
Guilty	Proven Guilty Not a thief (α)	Proven Guilty Really a thief

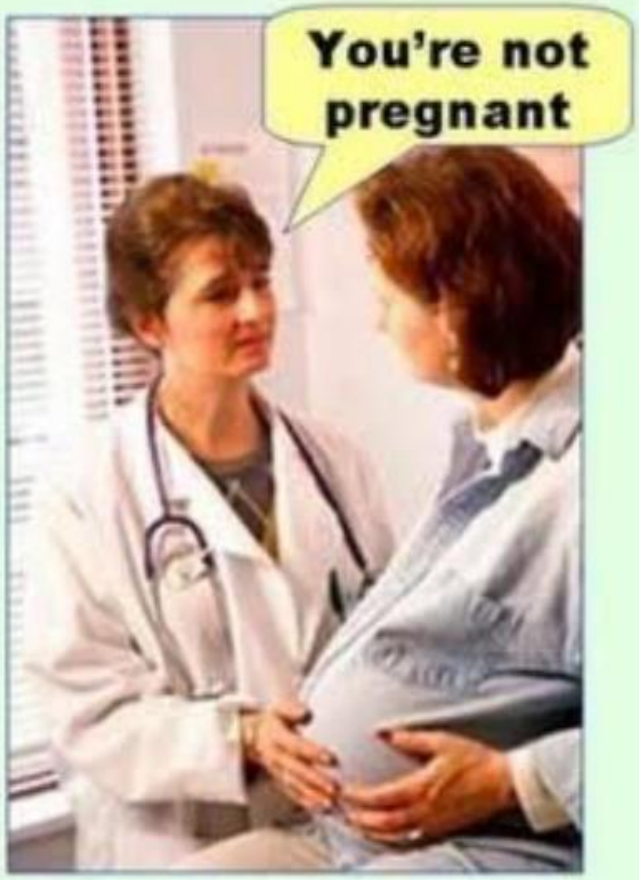
Type I Error: α

Type II Error: β

Type I error
(false positive)



Type II error
(false negative)



Terms in Statistics

	H_0 Correct	H_0 False
Accept H_0	$1 - \alpha$ Confidence Interval	β Error Type II
Reject H_0	α Error Type I	$1 - \beta$ Analysis Power

α: How big degree of type error I we can accept (1%, 5%, 10%) -> Wrong evidence

β: How big degree of type error II we can accept (10%, 15%, 20%) -> Lack of evidence

Metrics to measure error risk

Metrics to measure error risk to reject H_0 (Error Type I) is called p-value (Probability Value).

P-value is probability /chance that can represent **accepting H_1 or Rejecting H_0.**

For example:

Chance of him being a thief is 1%.

How to decide

There are 2 decisions:

- If P-Value < α, then we can accept H_1
- If P-Value > α, then we can accept H_0

Usually α is 1%, 5%, 10%

It depends on business decision or how confident the analysis are.

Example

Based on evidence provided, p-value of him being a thief is 0.03 (3%).

Standard α: 5%

What is the verdict of him being a thief?

DigitalSkola
Uncover The World Of Digital Skills With Us

Statistical Testing Methods

Statistical Methods

1. **T-test**
2. **Chi-square**
3. ANOVA
4. Pearson Correlation Test
5. Etc.

T-test

T-test is a statistics method that uses sample's average and distribution to compare between 2 population.

Type of T-test

Test for 1 population -> H_0: avg = 10 vs H_1: avg != 10

Test for 2 Independent population -> H_0: $avg_1 = avg_2$ vs H_1: $avg_1 != avg_2$

Assumptions

- Sample is normal distribution
- Or huge number of sample (central limit theorem)

Diet Plan

We want to prove which diet plan works best (alpha 5%)

Non-supplement Diet

Duration: **3 months**

Avg weight drop: **8.1 kg**

St Dev: **2.3 kg**

Sample: **30 users**

Supplement Diet

Duration: **3 months**

Avg weight drop: **8.5 kg**

St Dev: **1.5 kg**

Sample: **30 users**

Hypothesis:

H_0: avg = avg_2

H_1: avg_1 != avg_2

DigitalSkola
Uncover The World Of Digital Skills With Us

Let's test it here!

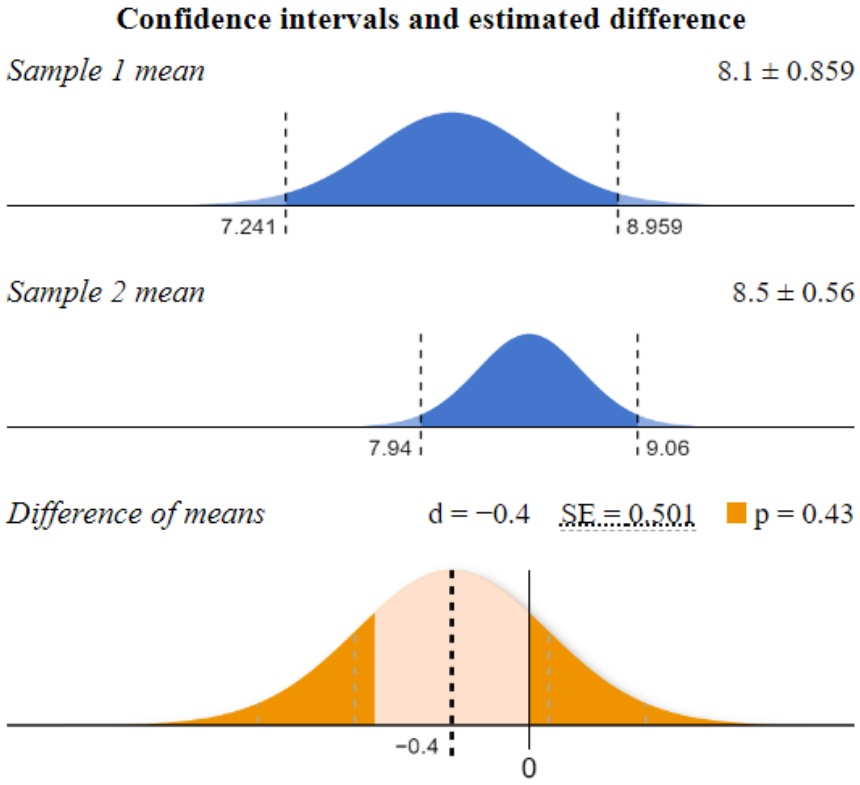
Test Result

DigitalSkola
Uncover The World Of Digital Skills With Us

Hypothesis:

H_0: average weight drop without supplement is **same** as weight drop with supplement

H_1: average weight drop without supplement is **different** from weight drop with supplement



Result

P-value = 43% and Alpha = 5%

P-value > Alpha: **H_0 Accepted / H_1 Rejected**

Verdict: average weight drop without supplement is **same** from weight drop with supplement

Meaning: There is **no different** between diet with or without supplement.

Chi-square

Chi-square Test is a method that is used to test if there is any relationship between two categorical variables.

It is also used to investigate whether distributions of categorical variables differ from one another.

Hypothesis

H_0: X and Y are independent.

H_1: X and Y are dependent.

Gender trouble maker

High School Teacher want to prove that gender is correlated with trouble (alpha 5%)

Boys

Students: **70**

Trouble maker: **25**

Normal Student: **45**

Girls

Students: **64**

Trouble maker: **10**

Normal Student: **54**

H_0: student's gender and trouble status is **independently correlated.**

H_1: student's gender and trouble status is **dependently correlated.**

DigitalSkola
Uncover The World Of Digital Skills With Us

Let's test it here!

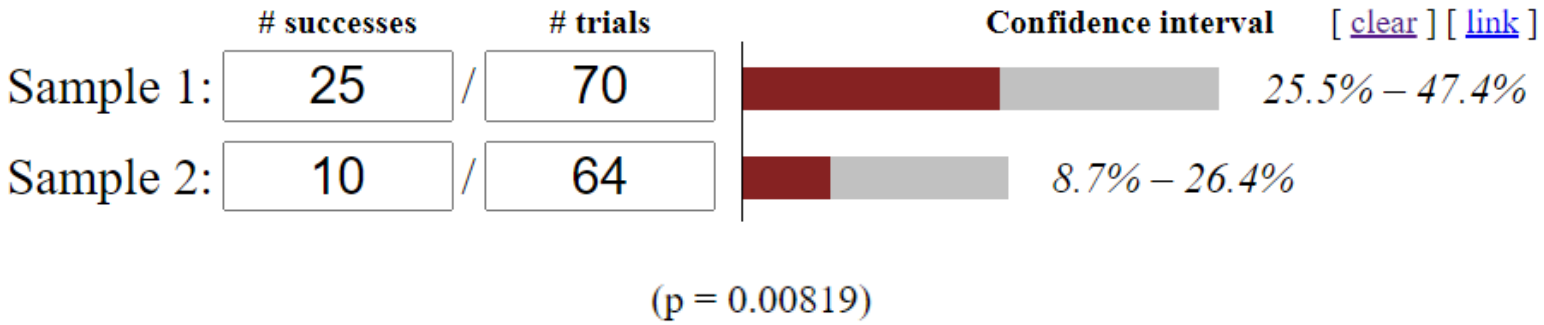
Test Result

DigitalSkola
Uncover The World Of Digital Skills With Us

Hypothesis:

H0: student's gender and trouble status is independently correlated.

H1: student's gender and trouble status is dependently correlated.



Result

P-value = 0.8% and Alpha = 5%

P-value < Alpha: **H_0 Rejected / H_1 Accepted**

Verdict: **There is correlation** between High school student's gender and trouble status.

Meaning: High school boys **tend to make trouble** than girls.

DigitalSkola
Uncover The World Of Digital Skills With Us

A/B Testing

Step to do experiment

1. Define an Experiment
2. Define Metrics
3. Define Duration of the experiment
4. Power Analysis
5. Post-Analysis

Define the Experiment

- What is the experiment name?

  AB Test of New Design for Obama Homepage

- Define the hypothesis?

  New design will improve the conversion rate

  H0: New and Existing design have **same** conversion rate

  H1: New and Existing design have **different** conversion rate

- Who is the participant?

  American people who visits the website

- What variables are going to test?

  Existing Design and New Design

Define The Metrics

- **North star Metric:**

  Metrics that is represent the grand vision.

  Obama is chosen as a president.

- **Primary Metrics:**

  AB Test main metrics. It is the decision metrics of experiment's success

  Conversion Rate of the website

- **Secondary Metrics:**

  Metrics that is used for complement the test.

  Duration of watched video

Define The Duration

- Usually AB Test Requires 14 days of experiment duration.
- We can shorten it the duration. However, shorter duration needs more respondents.

Power Analysis

- **Baseline rate:**

  Existing Value of primary metrics.

  based on this case is 4.9%

- **Minimum Detectable Effect (MDE):**

  It’s a minimum improvement over the conversion rate of the existing asset (baseline conversion rate) that you want the experiment to detect.

- **Sample Size:**

  Minimum sample size to validate the statistics test. Proportion between two samples must be balanced or 50:50.

Post-Analysis

DigitalSkola
Uncover The World Of Digital Skills With Us

We can use correct statistical methods to prove the AB Test. (T-test, Chi-square, Anova, etc.)

Let's say from the test we did:

Visitors of Existing	Converted Users	Visitors of New	Converted Users
41231	1855	42898	2445
53312	2666	55898	3689
32124	1638	36648	1942
34232	1677	31875	1530
23435	1265	32658	2155
65234	2805	68689	4190
43432	1781	40787	2243
12333	728	11650	699
32786	1639	30758	1969
30878	1667	28888	1704
37678	2185	25998	1248
37875	1667	35989	1727
31860	1593	33788	2162
30897	1823	32346	1682

DigitalSkola
Uncover The World Of Digital Skills With Us

Let's test it here!

DigitalSkola

Uncover The World Of Digital Skills With Us

Thank YOU