DATA FOLKZ®
#CATAPULT DATA LEADERS

Inferential Statistics

Understanding Inferential Statistics

There are two main areas of inferential statistics:

- **Estimating parameters.** This means taking a statistic from your sample data (for example the sample mean) and using it to say something about a population parameter (i.e. the population mean).

- **Hypothesis tests.** This is where you can use sample data to answer research questions. For example, you might be interested in knowing if a new cancer drug is effective. Or if breakfast helps children perform better in schools.

Descriptive and Inferential Statistics

BASIS FOR COMPARISON	DESCRIPTIVE STATISTICS	INFERENTIAL STATISTICS
Meaning	Descriptive Statistics is that branch of statistics which is concerned with describing the population under study.	Inferential Statistics is a type of statistics, that focuses on drawing conclusions about the population, on the basis of sample analysis and observation.
What it does?	Organize, analyze and present data in a meaningful way.	Compares, test and predicts data.
Form of final Result	Charts, Graphs and Tables	Probability
Usage	To describe a situation.	To explain the chances of occurrence of an event.
Function	It explains the data, which is already known, to summarize sample.	It attempts to reach the conclusion to learn about the population, that extends beyond the data available.

Example 1

Suppose we are interested in the exam marks of all the students in India. But it is not feasible to measure the exam marks of all the students in India. So now we will measure the marks of a smaller sample of students, for example 1000 students. This sample will now represent the large population of Indian students. We would consider this sample for our statistical study for studying the population from which it’s deduced.

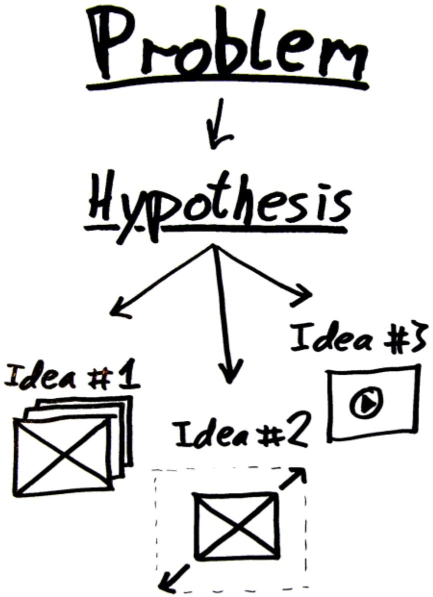
Hypothesis Testing

- A statistical hypothesis is an assumption about a population. This assumption may or may not be true.
- Hypothesis testing refers to the formal procedures used by statisticians to accept or reject statistical hypotheses made .
- In simple words , Hypothesis testing is used to test the validity of a claim that is made about a population using sample data.

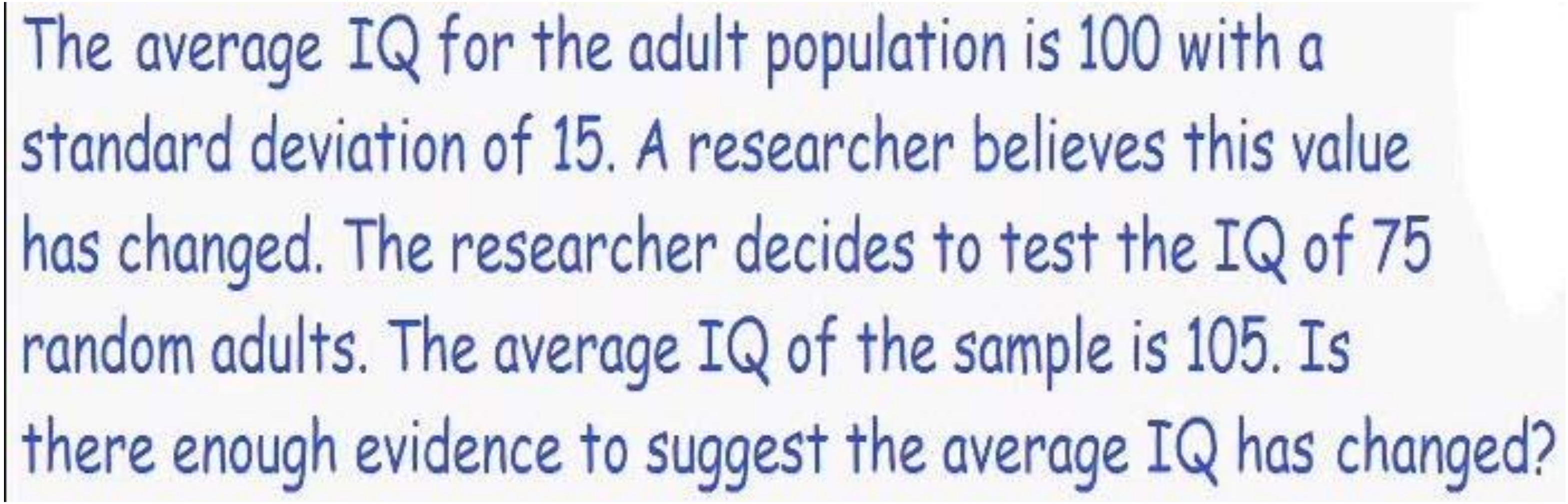
DATA FOLKZ®
#CATAPULT DATA LEADERS

What do you mean by Hypothesis ?

A supposition or proposed explanation made based on limited evidence as a starting point for further investigation.

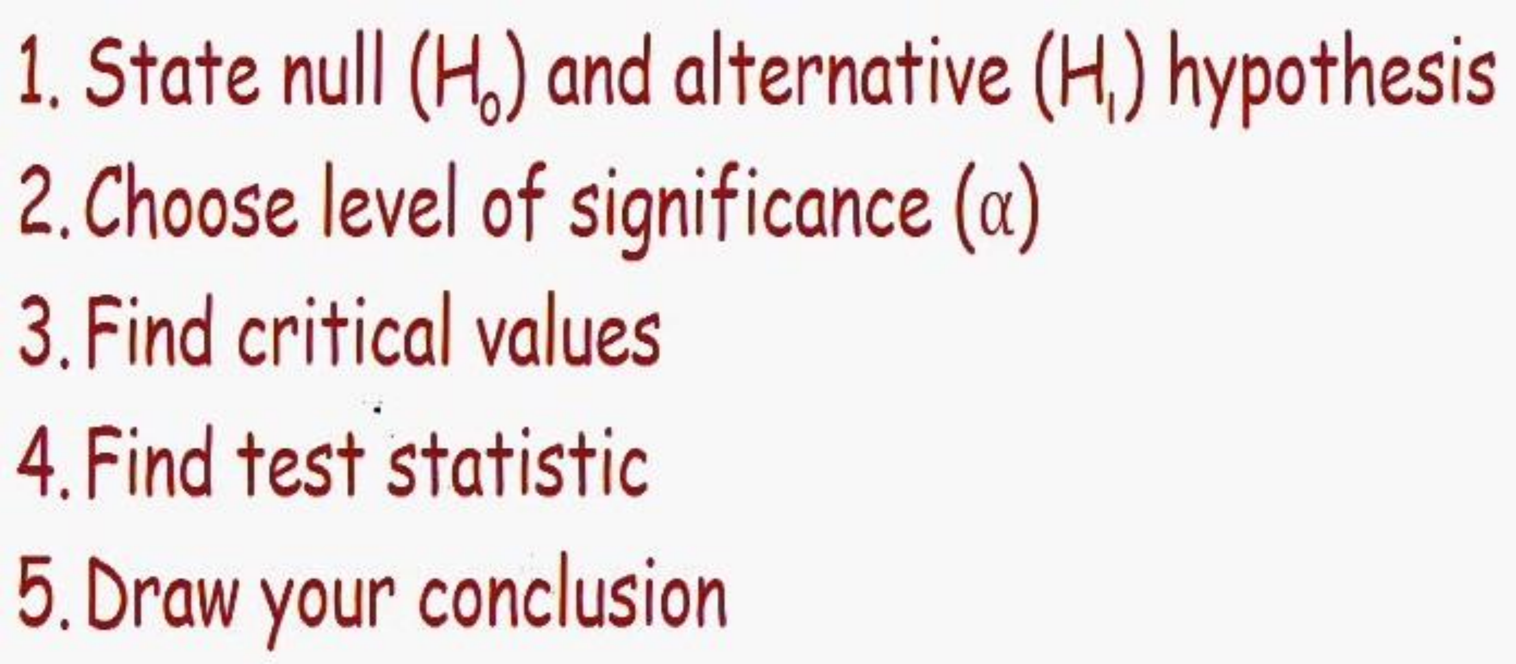


Let's understand Hypothesis Testing with an example



The average IQ for the adult population is 100 with a standard deviation of 15. A researcher believes this value has changed. The researcher decides to test the IQ of 75 random adults. The average IQ of the sample is 105. Is there enough evidence to suggest the average IQ has changed?

Hypothesis Testing Steps



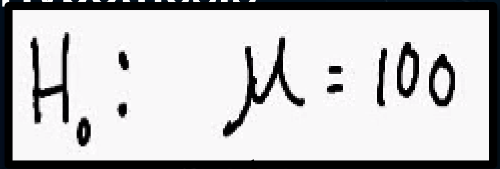
1. State null (H_0) and alternative (H_1) hypothesis
2. Choose level of significance (α)
3. Find critical values
4. Find test statistic
5. Draw your conclusion

Null Hypothesis
and
Alternate Hypothesis

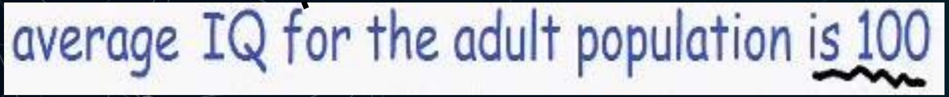
Null hypothesis (Ho) - It is the hypothesis which is being tested. It cannot be rejected without any strong evidence.

Alternative Hypothesis(Ha) - To every null hypothesis there exists an alternative hypothesis which holds true when null hypothesis is rejected.

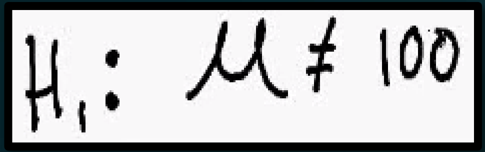
Step 1 : State Null & Alternative Hypothesis



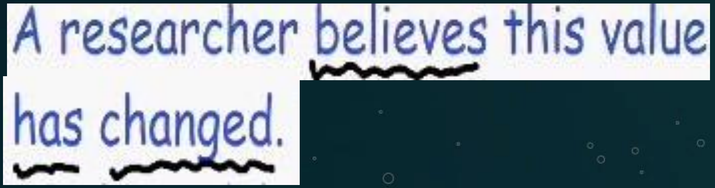
$H_0 : \mu = 100$



average IQ for the adult population is 100



$H_1 : \mu \neq 100$



A researcher believes this value has changed.

Simple Rules :

1. Null Hypothesis should always have = symbol
2. Alternative Hypothesis could have one of these > , < , >= , <= , !=
3. Alternative Hypothesis is **“what is being claimed”**
4. **When H_1 has != then we do a two-tailed test**
5. **When H_1 has either > , < , >=, <= then we do a one-tailed test**

Note : Hypothesis Testing is done to prove whether H_0 can be Rejected or Not be Rejected

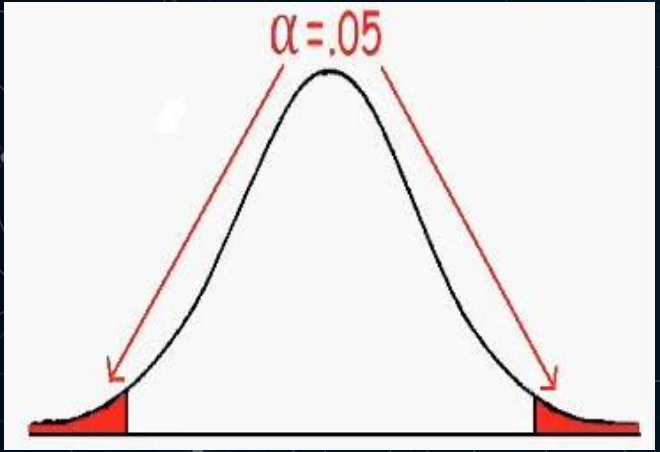
Step 2 : Choose level of Significance : A **significance level** , also known as alpha or α, is an evidentiary standard that a researcher sets before the study. It defines how strongly the sample evidence must contradict the null hypothesis before you can reject the null hypothesis for the entire population .

This is the area under the curve (marked red).

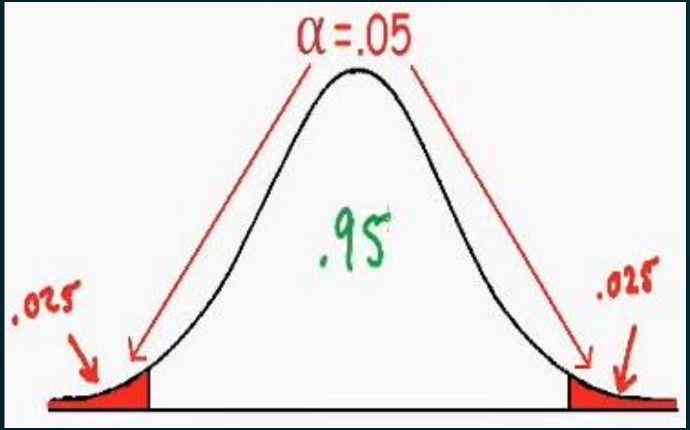
Usual values are 5% or 1%

5 % => .05

1 % => .01



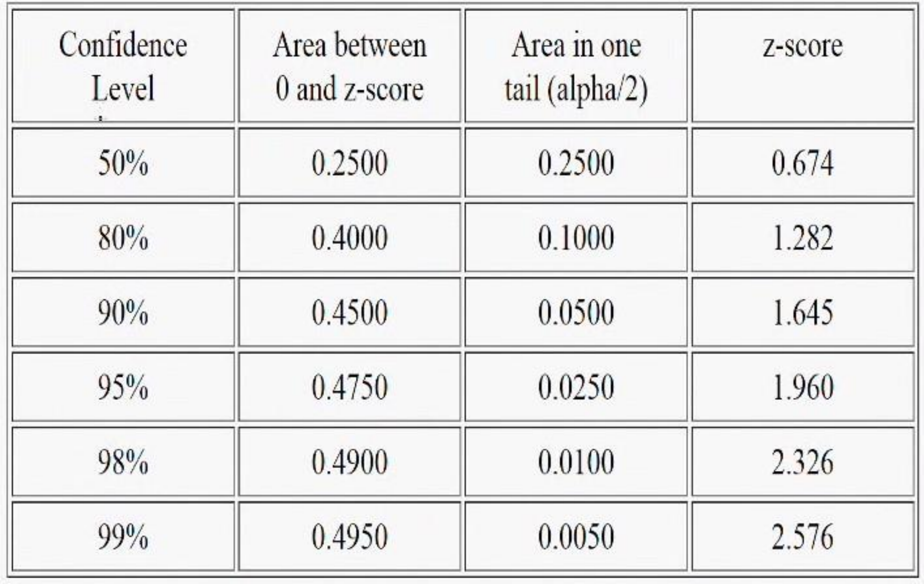
Note : If our **test statistic** comes in (red) **tail part** then we would **Reject H_0 else** we **will not Reject H_0**



Step 3 : Find Critical Values

Here we need find z-value or t-value.

Z-table :



Confidence Level	Area between 0 and z-score	Area in one tail (alpha/2)	z-score
50%	0.2500	0.2500	0.674
80%	0.4000	0.1000	1.282
90%	0.4500	0.0500	1.645
95%	0.4750	0.0250	1.960
98%	0.4900	0.0100	2.326
99%	0.4950	0.0050	2.576

How to decide between z-value or t-value ?

1. If the **std deviation of the population** is not given.

Or 2. If the **sample size < 30**

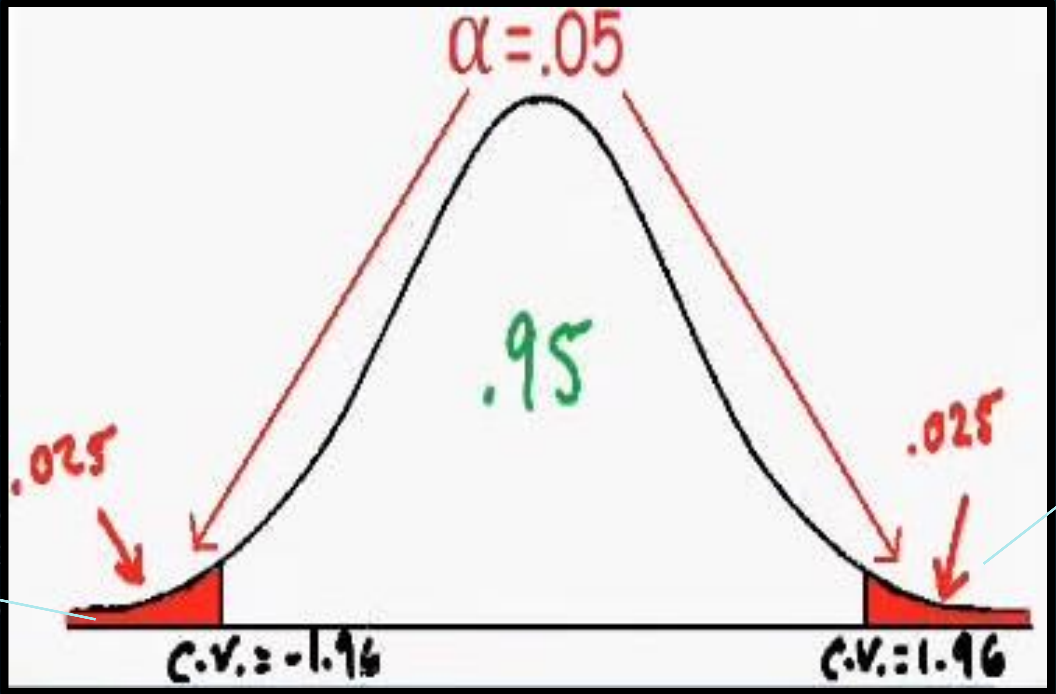
Then find a t-value , by using a t-table.

$\sigma = 15$

standard deviation of 15.

sample size > 30

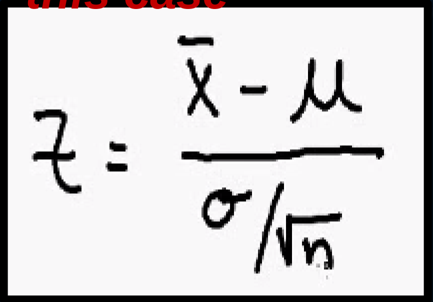
researcher decides to test the IQ of 75 random adults.



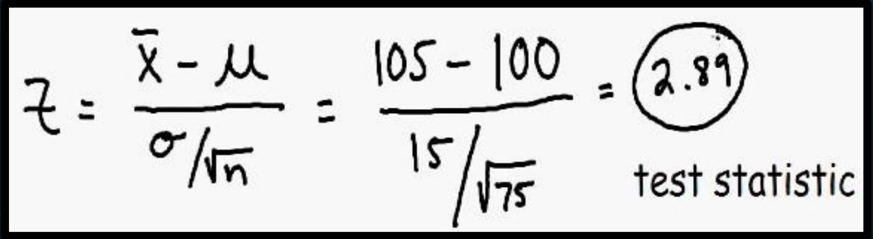
Rejection region

Rejection region

Step 4 : find test statistic. *i.e. the Z-value in this case*



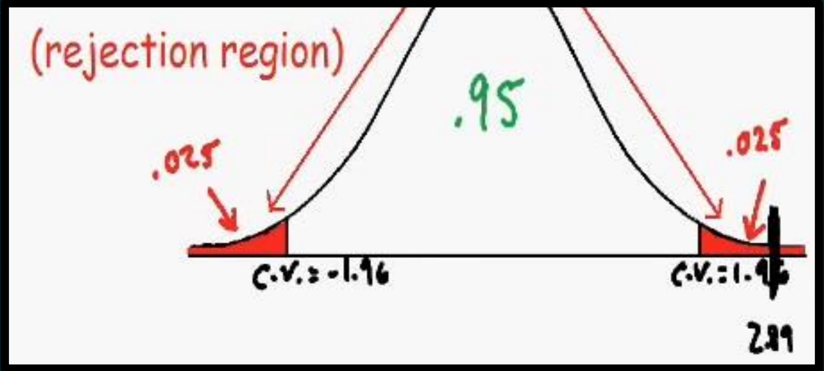
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$



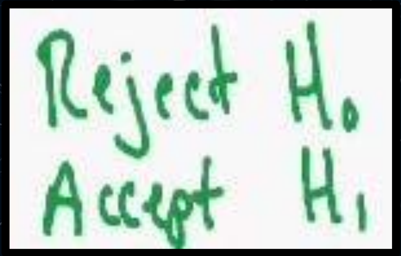
$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{105 - 100}{15/\sqrt{75}} = \boxed{2.89}$$

test statistic

So our test statistic value is **2.89**



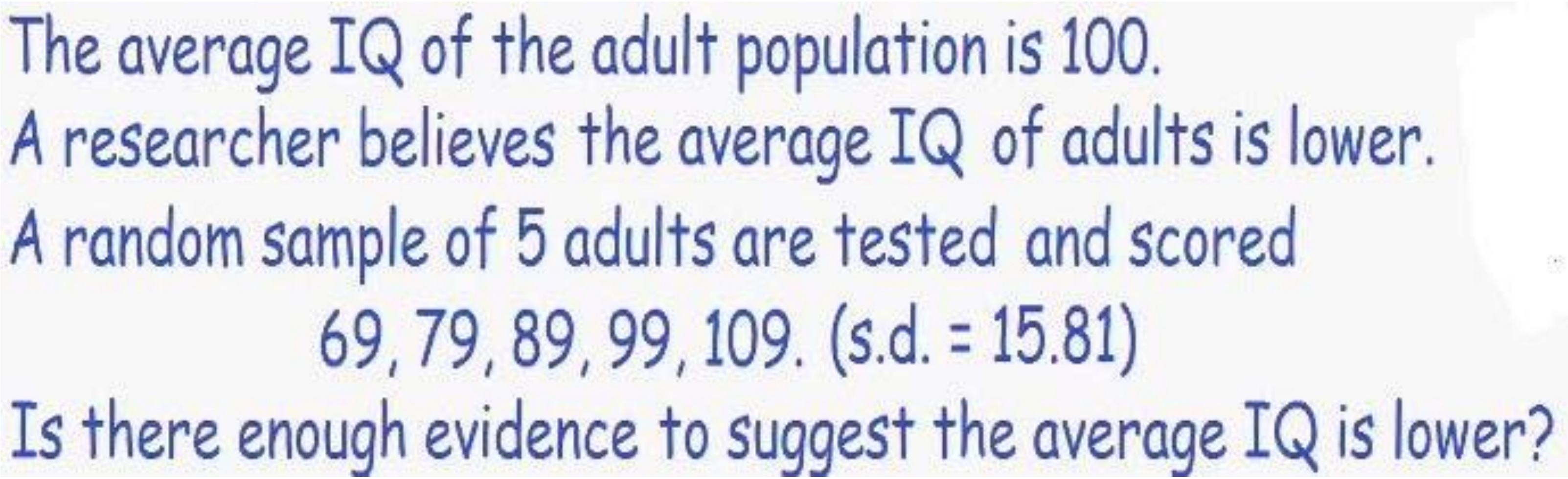
Step 5 : Conclusion



Reject H_0
Accept H_1

Hence, we say with 95 % confidence that there is enough evidence to support the claim : “Avg IQ has changed”

Another example of Hypothesis Testing



The average IQ of the adult population is 100.

A researcher believes the average IQ of adults is lower.

A random sample of 5 adults are tested and scored

69, 79, 89, 99, 109. (s.d. = 15.81)

Is there enough evidence to suggest the average IQ is lower?

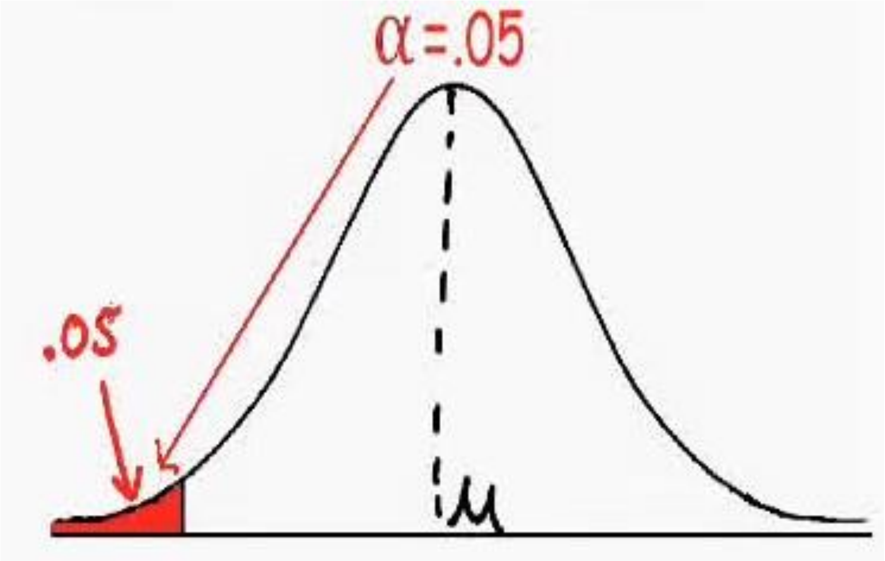
Solution :

Understanding the data given :

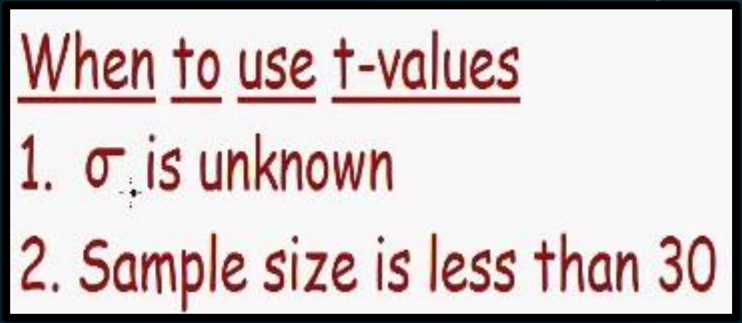
$H_0: \mu = 100$

$H_1: \mu < 100$

One-tailed test



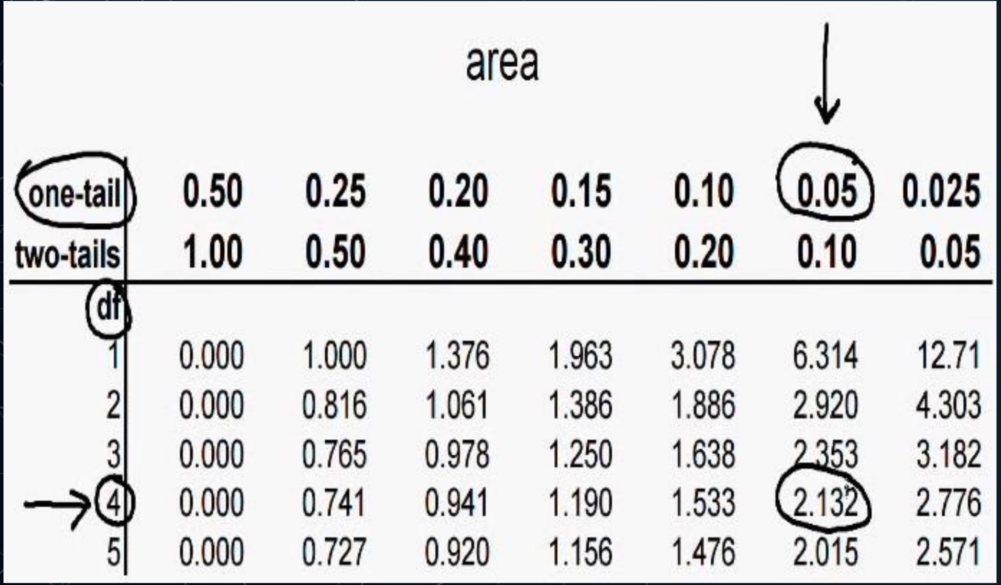
Here we need find z-value or t-value.



When to use t-values

1. σ is unknown
2. Sample size is less than 30

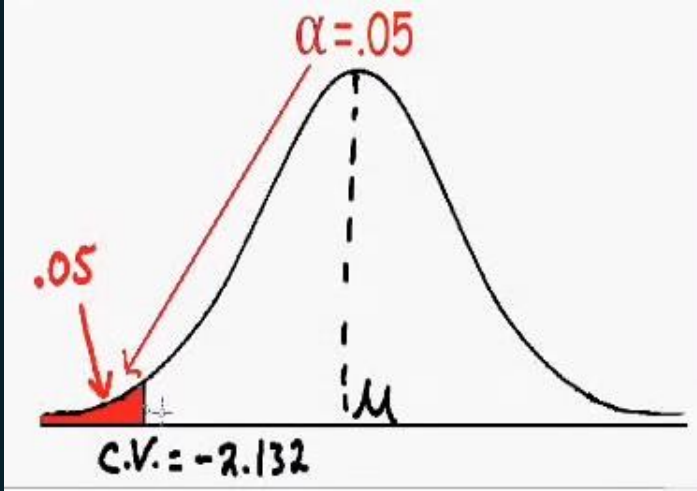
t-table



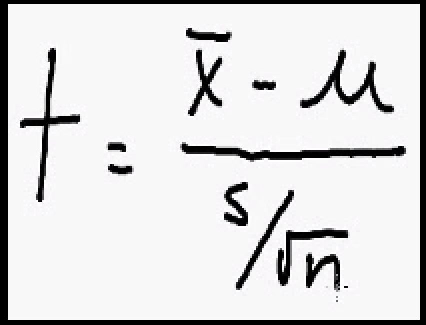
				area			
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05
df							
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571

df = degrees
of freedom
= (sampleSize - 1)

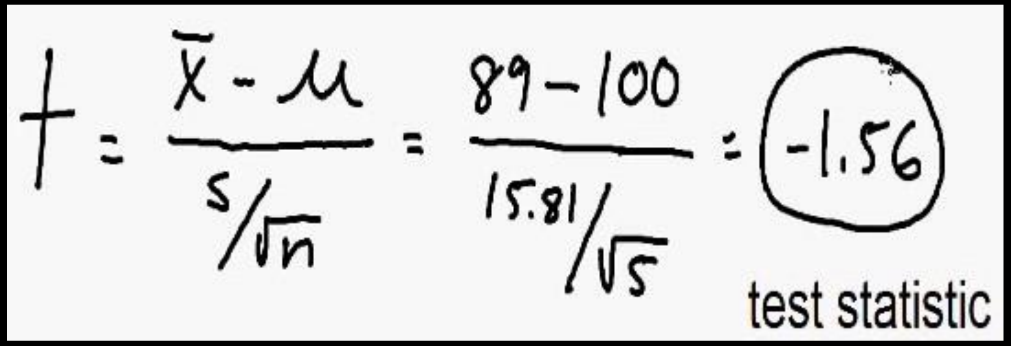
The **C.V** : Critical value on the left is **-ve** and the **C.V** on the Right is **+ve**



Find test statistic. *i.e. the t-value in this case*



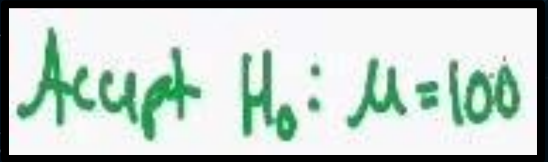
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$



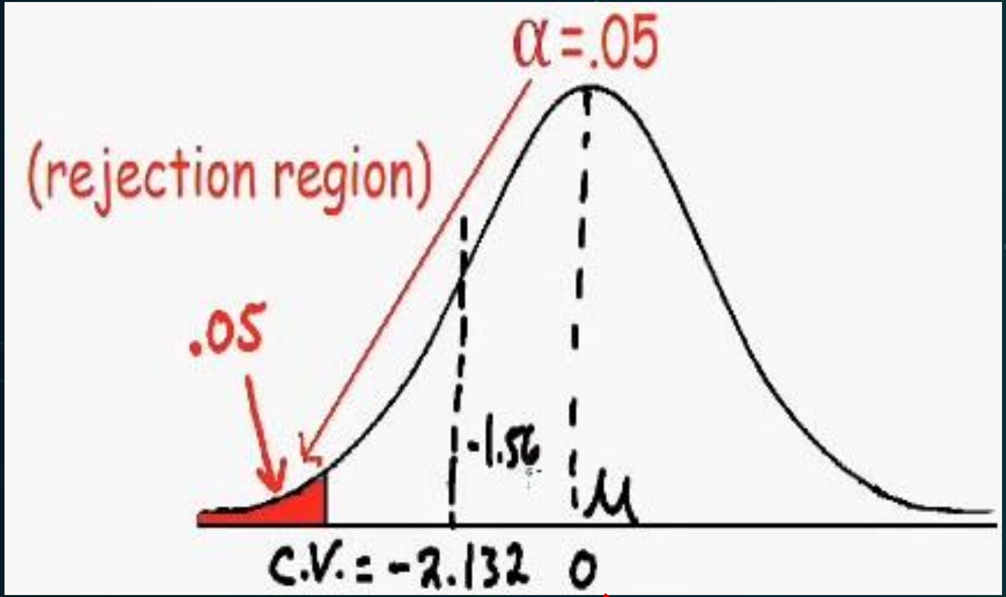
$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{89-100}{15.81/\sqrt{5}} = \boxed{-1.56}$$

test statistic

Conclusion :



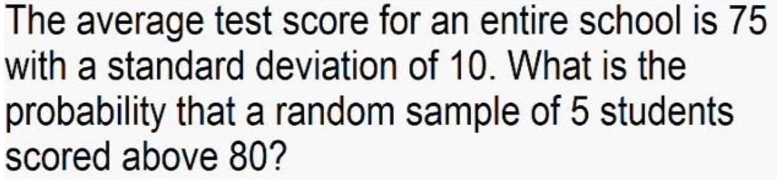
Accept $H_0 : \mu = 100$



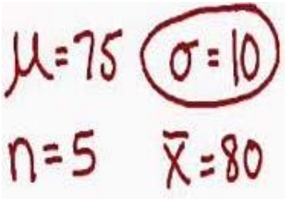
Hence, we say with 95 % confidence that there is enough evidence to support the claim : "Avg IQ = 100" or we lack evidence to support "Lower IQ's"

Center point of the Normal Distribution is always t = 0 or z=0

z-value vs. t-value calculation or z-statistic vs. t-statistic



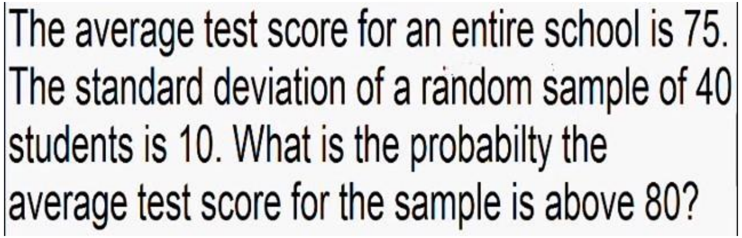
The average test score for an entire school is 75 with a standard deviation of 10. What is the probability that a random sample of 5 students scored above 80?



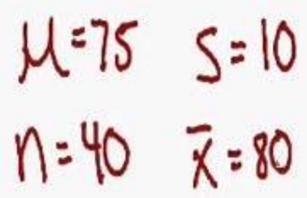
$\mu = 75 \quad \boxed{\sigma = 10}$

$n = 5 \quad \bar{x} = 80$

Hence we use Z test



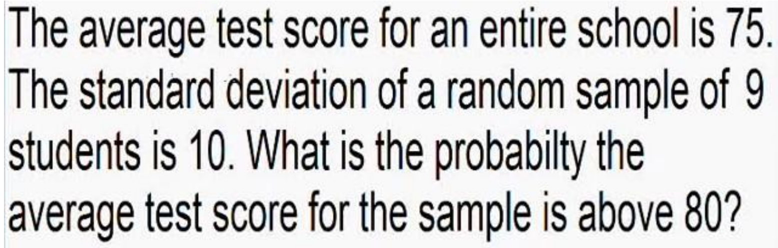
The average test score for an entire school is 75. The standard deviation of a random sample of 40 students is 10. What is the probabilty the average test score for the sample is above 80?



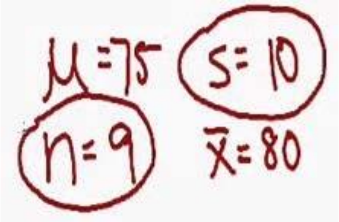
$\mu = 75 \quad s = 10$

$n = 40 \quad \bar{x} = 80$

Hence we use Z test



The average test score for an entire school is 75. The standard deviation of a random sample of 9 students is 10. What is the probabilty the average test score for the sample is above 80?



$\mu = 75 \quad \boxed{s = 10}$

$\boxed{n = 9} \quad \bar{x} = 80$

Both conditions met. Hence use t-test

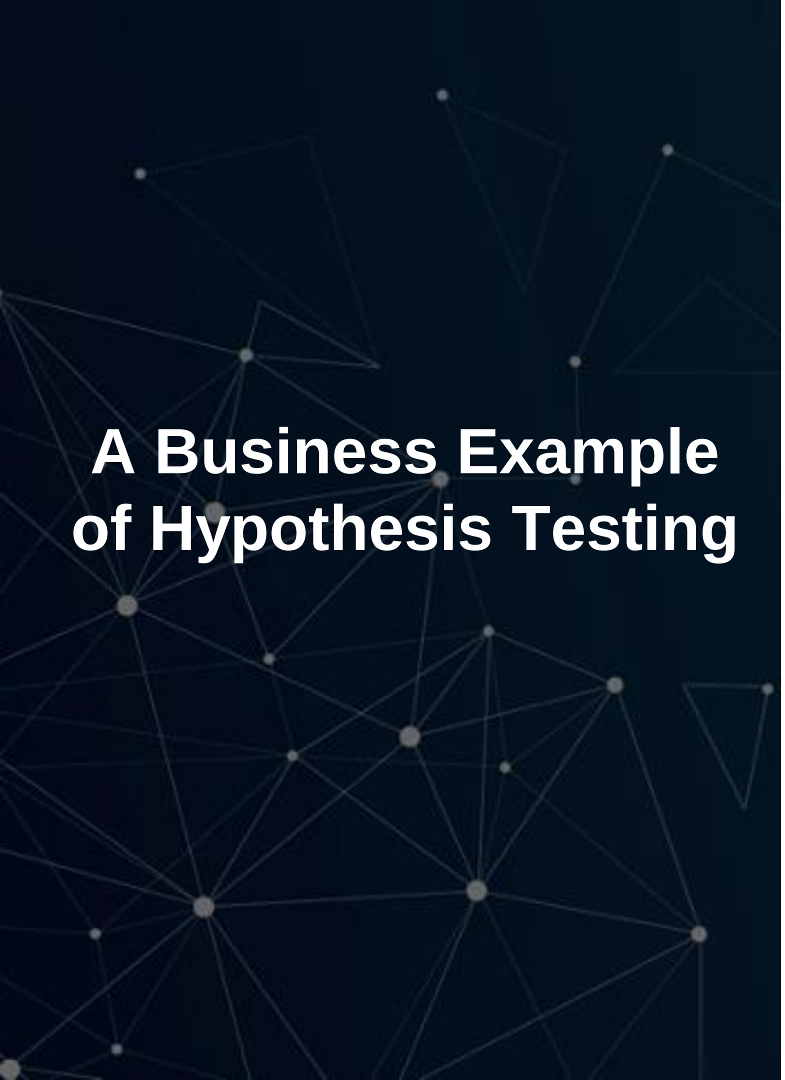
DATA FOLKZ®
#CATAPULT DATA LEADERS

Why You Should Perform Hypothesis Testing and Statistical Tests ?

- Hypothesis testing is a form of inferential statistics that allows us to draw conclusions about an entire population based on a representative sample.

- Using Hypothesis testing, valuable business decisions could be made in no time and with precision. This method has helped many pharmaceutical drugs and medical procedures in testing.

- Hypothesis Tests, or Statistical Hypothesis Testing, is a technique used to compare two datasets, or a sample from a dataset. It is a **statistical inference method** so, in the end of the test, you'll **draw a conclusion -** you'll infer something - about the characteristics of what you're comparing .



A Business Example of Hypothesis Testing

Let's take quality control for example. Say Widgets R Us is manufacturing a widget that must have a width of 150 mm, with a small tolerance. In this case hypothesis testing can be used . Here , Basic hypotheses might be:

H_0: The widget sizes equal 150
H_A: The widget sizes do not equal 150.

For the statistics used in the test, Widgets R Us can randomly sample and measure the average widget size over 30 production runs. If they want a fairly
strict assurance that the samples are close to the required value, they can use a 5
(0.05) significance level for evaluation .

That's how Hypothesis testing comes into picture in the business problem .

Errors in Hypothesis Testing

Two types of error in decision making:

- Type I ($\propto$) – Falsely reject H_0
- Type II (β) – Fail to reject H_0 when it is false

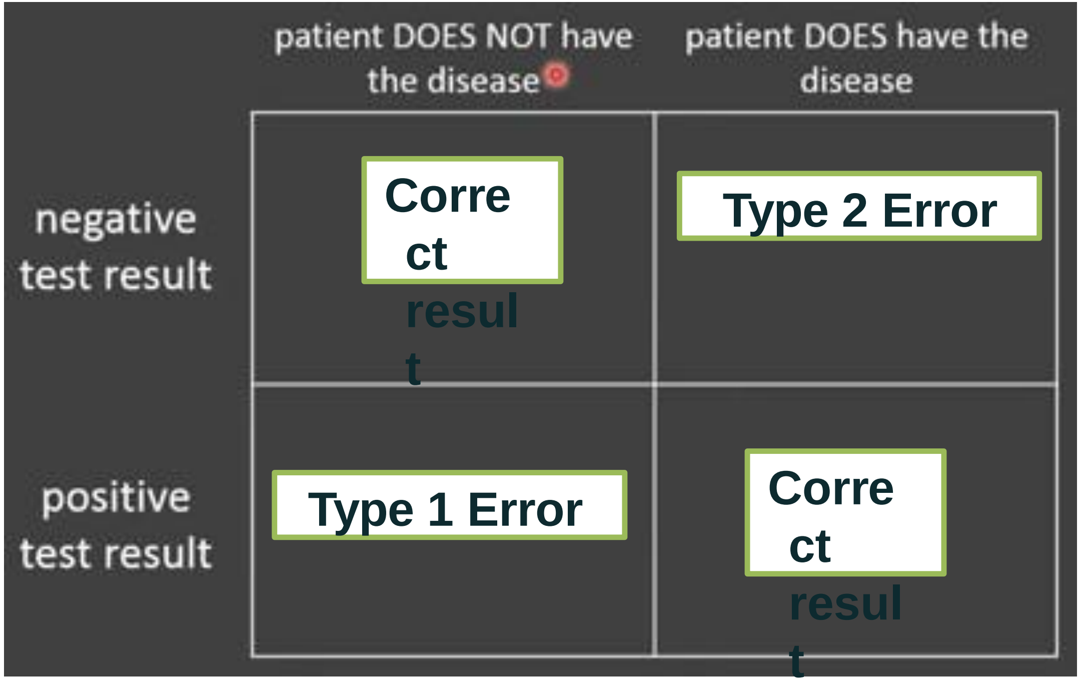
- $\propto$ and β are inversely proportional, meaning that decreasing one will increase the other.

- Typically, the Type I error rate is set to a moderate level, resulting in a reasonable Type II error rate.

Example

Truth about the patient

Test Results



	patient DOES NOT have the disease	patient DOES have the disease
negative test result	Correct result	Type 2 Error
positive test result	Type 1 Error	Correct result

Central Limit Theorem

- The central limit theorem (CLT) states that the distribution of sample means approximates a normal distribution as the sample size gets larger. Sample sizes equal to or greater than 30 are considered sufficient for the CLT to hold.
- In simple words , when you have roughly 30 or more observations in your sample, the average of those numbers is part of a bell-shaped curve.
- A key aspect of CLT is that the average of the sample means, and standard deviations will equal the population mean and standard deviation.
- A sufficiently large sample size can predict the characteristics of a population accurately.

Assumptions of Central Limit Theorem

DATA FOLKZ®
#CATAPULT DATA LEADERS

1. The Mean of Sample is equal to the Mean of Population.

2. The standard deviation of Sample is (S.D. of population/$\sqrt{n}$).

3. Standard Error is the difference between S.D. of Population and S.D. of Sample

Central Limit Theorem

Central limit theorem is applicable for a sufficiently large sample sizes (n≥30). The formula for central limit theorem can be stated as follows:

$$\mu_{\bar{x}} = \mu \qquad \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

Where,
μ = Population mean
σ = Population standard deviation
$\mu_{\bar{x}}$ = Sample mean
$\sigma_{\bar{x}}$ = Sample standard deviation
n = Sample size

Central Limit Theorem Example

Q The record of weights of male population follows normal distribution. Its mean and standard deviation are 70 kg and 15 kg respectively. If a researcher considers the records of 50 males, then what would be the mean and standard deviation of the chosen sample?

Ans.
Mean of the population μ = 70 kg
Standard deviation of the population = 15 kg
sample size n = 50

Mean of the sample is given by: $\mu_{\bar{x}}$ = 70 kg

Standard deviation of the sample is given by:

$$\sigma_{\bar{x}} = {}^{\sigma}/_{\sqrt{n}}$$

$$\sigma_{\bar{x}} = {}^{15}/_{\sqrt{50}}$$

$\sigma_{\bar{x}}$ = 2.1 kg

P-Value

A P-value measures the strength of evidence in support of a null hypothesis .

When you perform a statistical test a *p*-value helps you determine the significance of your results in relation to the null hypothesis.

If the p value is less than significance value (0.05) then we reject null hypothesis and if the p value is greater than significance value (0.05) then we fail to reject null hypothesis .

For example, say that a fair coin is tested for fairness (the null hypothesis). At a significance level of 0.05, the fair coin would be expected to (incorrectly) reject the null hypothesis in about 1 out of every 20 tests.

Importance of p-

Perform this experiment :

1. flip a coin -> assume we get "head"
   P(H) = 0.5
2. flip the coin again -> assume we get "head" again
   P(H | H) = 0.5 x 0.5 = 0.25

Now before proceeding : "Is this a fair coin ?" [What's your answer Yes / No ?]

[What-ever your ans ? How do you support your answer ?]

Here Null hypothesis = H0 = Is a fair coin
Lets keep on tossing our coin to figure out whether we should :
1. Reject H0
2. or Not Reject H0
3. flip the coin -> assume we get "head" again P(H | 2H) = 0.5 x 0.5 x 0.5 = 0.125

4. flip the coin -> assume we get "head" again

$P(H \mid 3H) = 0.5 \times 0.5 \times 0.5 \times 0.5 = 0.0625$

That’s 6% chance of we getting 4 heads in a row.

5. {last time} flip the coin -> assume we get "head" again

$P(H \mid 4H) = 0.5 \times 0.5 \times 0.5 \times 0.5 \times 0.5 = 0.0312$

That’s 3% chance of we getting 5 heads in a row.

p-value = 3%

We can say that chances of we getting a 5H in a row is just 3% or not getting 5H in a row is 97%.

Hence we can Reject our Null Hypothesis H0 i.e. the coin is fair.

[**Note** : When we reject H0 we accept H1 i.e. the Alternative Hypothesis i.e. the coin is not fair !!]

Thank You !