Statistics for Data Science: Inferential Statistics

Hari, Tanggal

Data Analytics

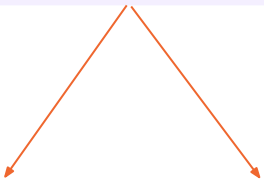
Program Zenius Studi Independen Bersertifikat Bersama Kampus Merdeka

1. Probability Mass Function and Probability Density Function
2. Hypothesis Testing Concept
3. T-test
4. Real Case Example

What is Inferential Statistics?

Inferential Statistics

Inferential statistics allows you to make predictions (“inferences”) from the data. With inferential statistics, you take data from samples and make generalizations about a population.



Estimating Parameters

Hypothesis Testing

Probability Mass Function and Probability Density Function

Discrete vs Continuous

Discrete Variable: variable that the variants are countable (we can count it).

Example:
Rolling Dice Numbers

Continuous Variable: variable that the values are all real numbers within a certain interval.

Example:
Height, Weight

Probability Mass Function

A function that gives the probability that a discrete random variable is exactly equal to some value

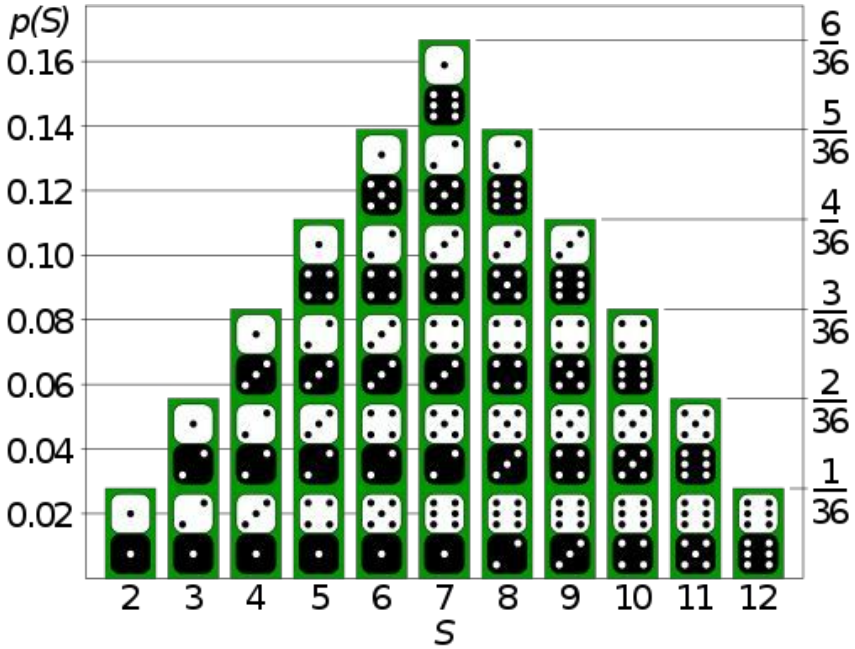
$$p_X(x) = \mathrm{P}(X = x), \ \forall x \in \mathbb{R}$$

Example: X is a random variable that represents the number/value in 1 rolling dice

Value	Probability
1	1/6
2	1/6
3	1/6
4	1/6
5	1/6
6	1/6

$$p_{\mathrm{X}}(x) = \begin{cases} 1/6 & \text{if } x = 1 \\ 1/6 & \text{if } x = 2 \\ 1/6 & \text{if } x = 3 \\ 1/6 & \text{if } x = 4 \\ 1/6 & \text{if } x = 5 \\ 1/6 & \text{if } x = 6 \\ 0 & \text{otherwise} \end{cases}$$

Probability Mass Function

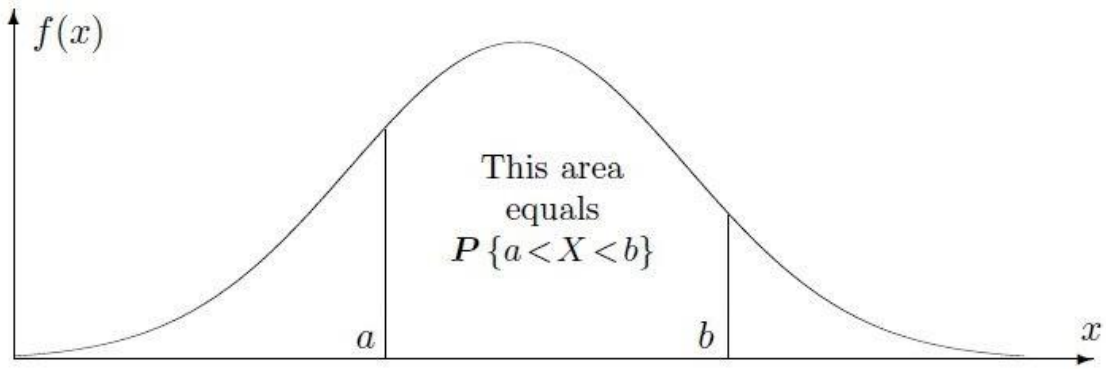


Rolling 2 dices:
Let X be a random variable that represents the sum when two dices are rolled.

x(sum)	2	3	4	5	6	7	8
$P(x)$	$\frac{1}{16}$	$\frac{1}{8}$	$\frac{3}{16}$	$\frac{1}{4}$	$\frac{3}{16}$	$\frac{1}{8}$	$\frac{1}{16}$

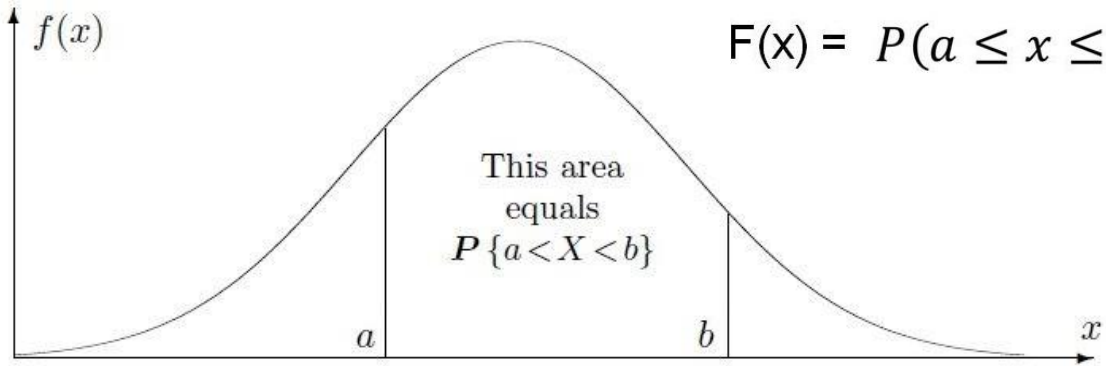
Probability Density Function

Define the random variable's probability coming within a distinct range of values.



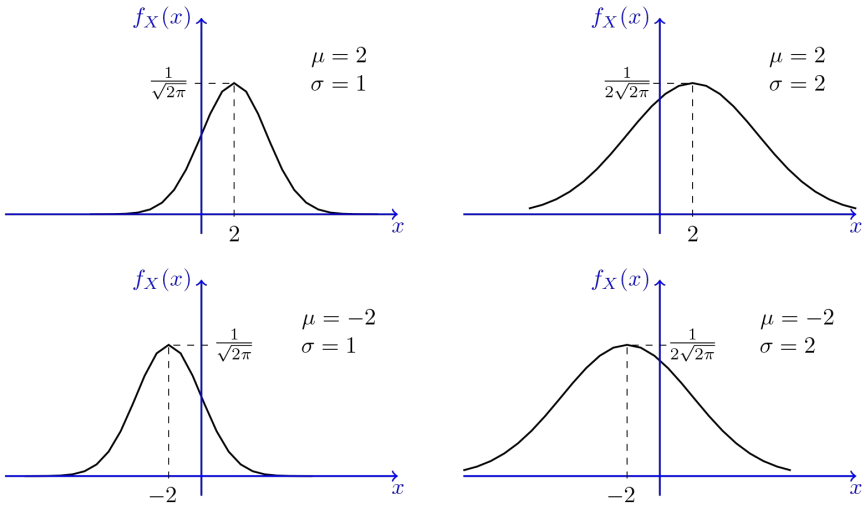
Probability Density Function

Finding the probability at certain interval is equal to calculating the area under the PDF curve.



$$F(x) = P(a \leq x \leq b) = \int_a^b f(x)dx \geq 0$$

Example: Normal Distribution PDF



$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

$$\text{F(x)} = P(a \leq x \leq b) = \int_a^b f(x)dx \geq 0$$

Pop Quiz!

If we toss a single coin, the probability of getting a head is?

A. 0.5

B. 0.25

C. 1

Pop Quiz!

The probability mass function is always less than or equal to 1.

A. True

B. False

Pop Quiz!

If we roll 1 dice, the probability of getting 7 is?

A. 1/6

B. 0

C. 2/6

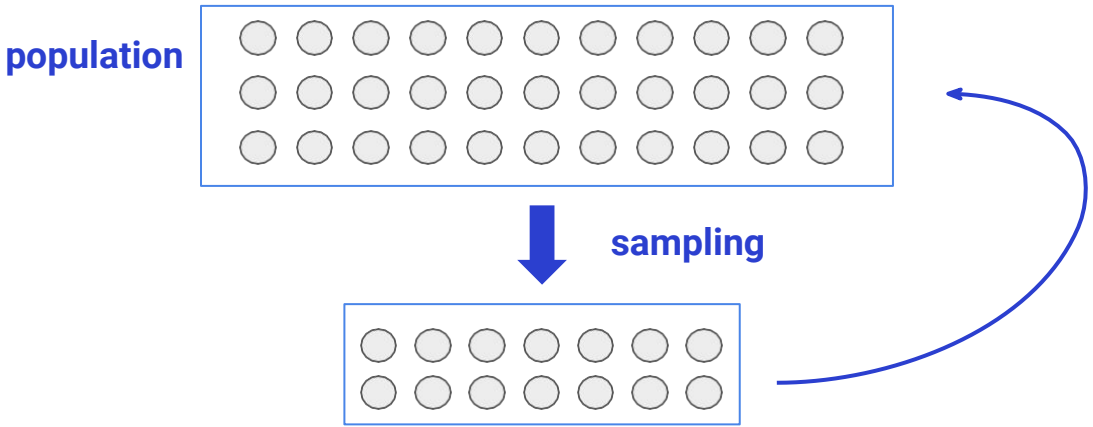
Hypothesis Testing Concept

What is Hypothesis Testing

Hypothesis testing is a **statistical method used to determine a possible conclusion** from two different, and likely conflicting, hypotheses.

It is tested on **a random sample** to **make a conclusion about a population**.

Hypothesis testing is widely used in business, marketing, manufacturing, clinical trials, and many more.



Examples

- Biology:
  - Determine whether some new treatments cause increased growth, stamina, or immunity of a plant/animal.
- Clinical Trials:
  - Determine whether some drugs cause improve outcome in patients.
- Marketing/Advertising
  - Determine the effect of certain campaign/promotion to users' behaviour/transactions
- Manufacturing
  - Determine if some new processes/techniques cause a change in the number of defective products produced

etc.

H0 and H1

In a hypothesis test, there are two hypotheses proposed: **a null hypothesis** and **an alternative hypothesis**

Null Hypothesis (H0)

The null hypothesis states that a population parameter (such as the mean, the standard deviation, and so on) **is equal to a hypothesized value**.

Alternative Hypothesis (H1)

The alternative hypothesis states that a population parameter **is smaller, greater, or different than the hypothesized value** in the null hypothesis.

Hypothesis testing purpose: <u>To test whether we can reject the null hypothesis (H0)</u>

H0 and H1

We have math exam results for a sample of students who took a training course for a national exam. We want to know if the mean of all students score above the national average of 850.

Null Hypothesis (H0)

The null hypothesis states that a population parameter (such as the mean, the standard deviation, and so on) **is equal to a hypothesized value**.

$$H0: \mu = 850$$

Alternative Hypothesis (H1)

The alternative hypothesis states that a population parameter **is smaller, greater, or different than the hypothesized value** in the null hypothesis.

$$H1: \mu > 850$$

Confidence Interval

Confidence interval is **the range of values within which the true population value would lie.**

Formula for a standard normal data:

$$\bar{X} - Z \cdot \frac{\sigma}{\sqrt{n}} \quad \textbf{to} \quad \bar{X} + Z \cdot \frac{\sigma}{\sqrt{n}}$$

Where **Z refers to the Z-value (standard normal distribution), depends on the level of confidence interval**. If we want to find 95% confidence interval, then we find the correspondent Z-score with the probability of 95% from the **Z table.**

Confidence Level

Definition

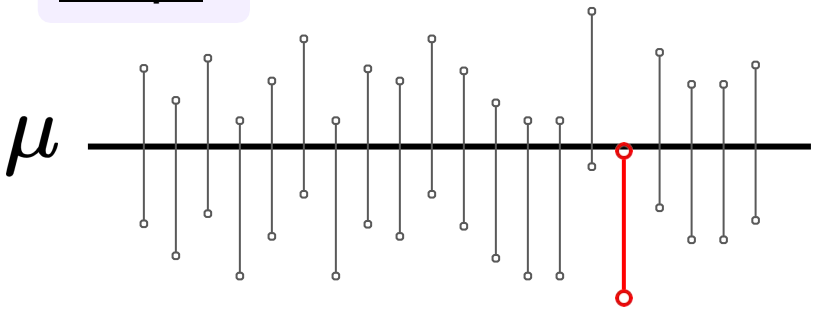
The confidence level represents the percentage of intervals that would include the population parameter if we took samples from the same population again and again.

Most frequently used confidence level: 95%

Meaning:
If we collected one hundred samples, and made one hundred 95% confidence intervals, we would expect **approximately 95 of the intervals to contain the population parameter**, such as **the mean of the population**.

Confidence Level

Example



- **The horizontal black line** represents the fixed value of the unknown population mean, μ.
- **The vertical grey confidence intervals** that overlap the horizontal line contain the value of the population mean.
- **The red confidence interval** is completely below the horizontal line.

A 95% confidence level indicates that **19 out of 20 samples (95%)** from the same population will produce confidence intervals that contain the population parameter.

Significance Level or Alpha

The significance level, also denoted as alpha or α, is the probability of rejecting the null hypothesis when it is true.

The formula of significance level is:

1 - confidence level

If we chose to use 95% of confidence level in our hypothesis test, then the **significance level is 5% or 0.05.**

Test Statistic

A test statistic is **a random variable that is calculated from sample data** and **used in a hypothesis test**. We can use test statistics to determine whether to reject the null hypothesis (H0).

A test statistic has different formula depends on the hypothesis test type.

Different hypothesis tests use different test statistics based on the probability model assumed in the null hypothesis. Common tests and their test statistics include:

Hypothesis test	Test statistic
Z-test	Z-statistic
t-test	t-statistic
ANOVA	F-statistic
Chi-square test	Chi-square statistic

P-Value

The p-value is a **measure of the strength of the evidence in your data against H0**.

Usually, the smaller the p-value, the stronger the sample evidence is for rejecting H0.
More specifically, the p-value is the smallest value of α that results in the rejection of H0.

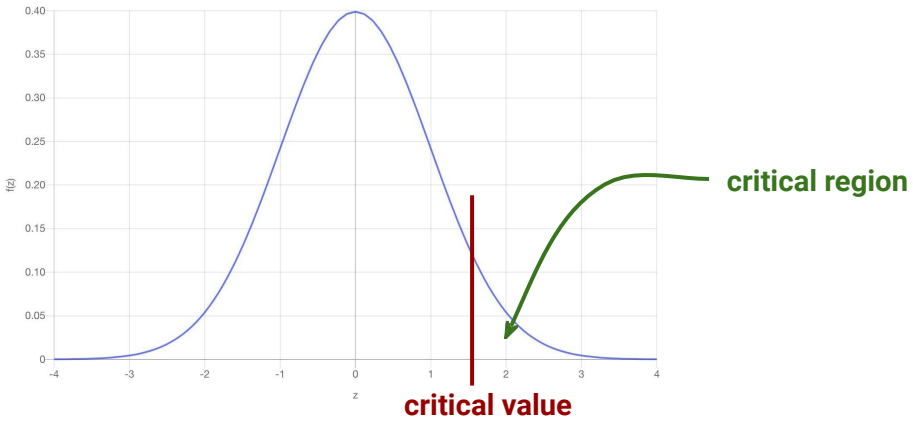
The p-value is calculated using the sampling distribution of the test statistic under the null hypothesis and the sample data.

Note:
It is usually calculated automatically in programming language libraries/statistical analysis tools like in Python, R, and SPSS.

Critical Value and Critical Region

A critical value is **a point on the distribution of the test statistic under the null hypothesis that defines a set of values that call for rejecting the null hypothesis**.

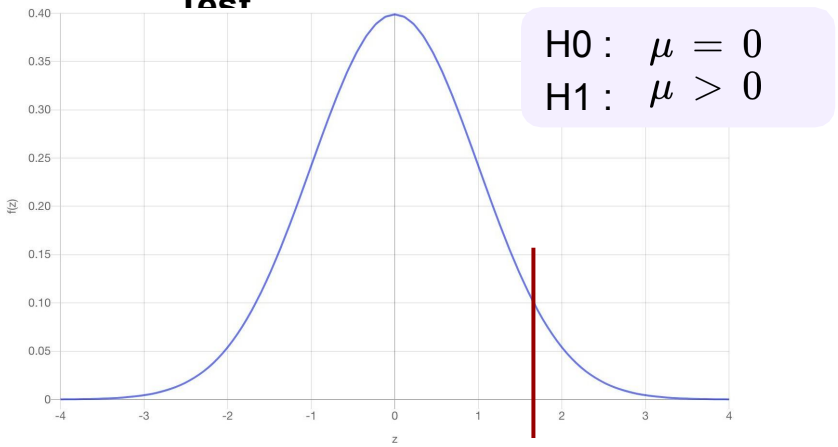
Critical value is obtained from **the correspondent score in the probability distribution table**, depending on the hypothesis test type.
(For example: If we run **t-test**, then **we can get critical value from the t table** using the chosen significance level)



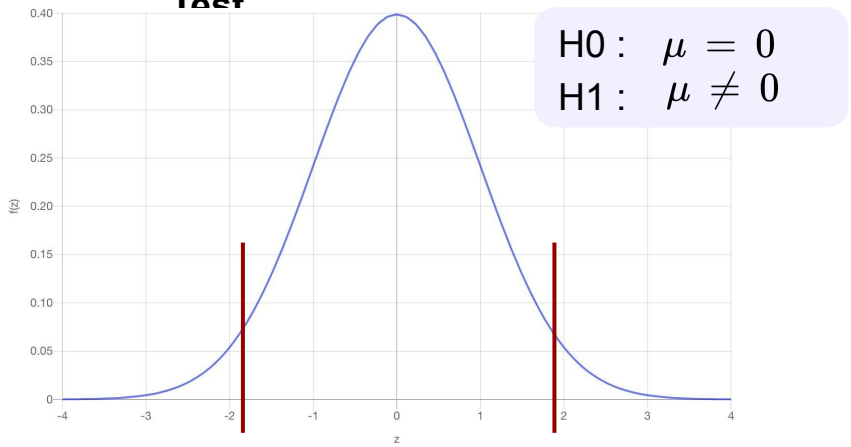
One Sided & Two Sided Test

There are two types of a hypothesis test based on the alternative hypothesis proposed.

One Sided/Tailed Test



Two Sided/Tailed Test



Deciding Whether to Reject H0

Using P-value and α

- If the p-value is less than or equal to α, we reject H0;
- If it is greater than α, we fail to reject H0

Using test statistic and critical region

- If the test statistic falls inside the critical region, we reject H0;
- If it falls outside the critical region, we fail to reject H0

Summary: Hypothesis Testing Steps

1. Specify the hypotheses (H0 and H1)
2. Choose a significance level (also called alpha or α)
3. Run the test and calculate test statistic and P-value
4. Find critical region
5. Compare P-value and significance level or test statistic and critical region
6. Decide whether to reject or fail to reject the null hypothesis

Pop Quiz!

What hypothesis states equality or no difference, or no relationship/effect?

A. H0

B. H1

Pop Quiz!

If the test is two-tailed, the critical region, with an area equal to α, will be on the left side of the mean.

A. True

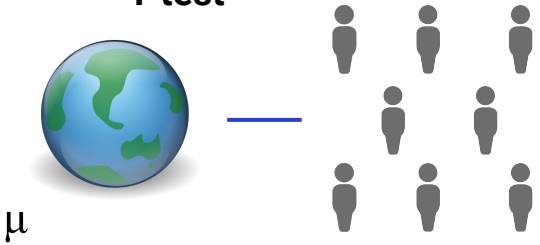
B. False

T-Test

What is T-Test?

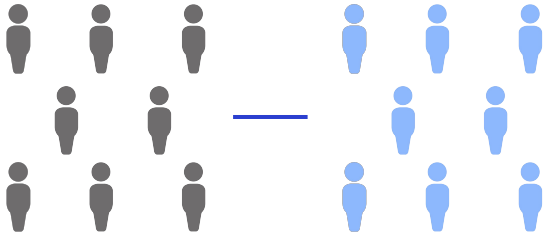
T-test is a statistical test that is used to compare **the means of two groups**. It is often used in hypothesis testing to determine whether **a process or treatment actually has an effect on the population of interest**, or **whether two groups are different from one another**.

One sample T-test



Is there a difference between a group and the population? Or is a group's mean equal to a certain value?

Two sample T-test

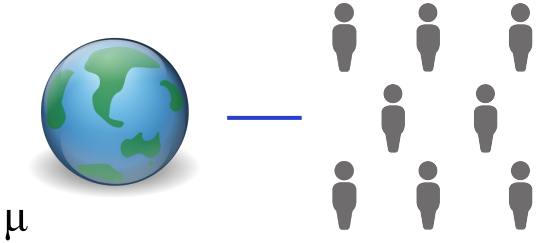


Is there a difference between two groups?

What is T-Test?

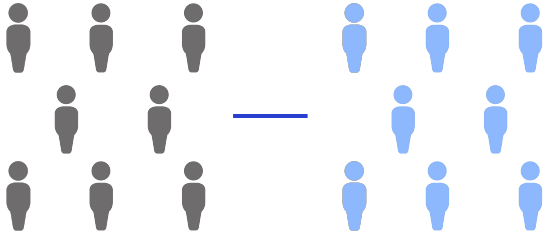
T-test is a statistical test that is used to compare **the means of two groups**. It is often used in hypothesis testing to determine whether **a process or treatment actually has an effect on the population of interest**, or **whether two groups are different from one another**.

One sample T-test



Is there a difference between a group and the population? Or is a group's mean equal to a certain value?

Two sample T-test



Is there a difference between two groups?

One Sample T-test

One sample T-Test compares the mean of our sample data to a known value.
For example, we might want to know how our sample mean compares to the population mean. This hypothesis test is used when we don't know the population standard deviation or we have a small sample size (t-distributed data).

Assumptions:

- Data is collected randomly (representative, randomly selected portion of the total population).
- The data is approximately normally distributed.
- The data is independent.

One Sample T-test Hypotheses

Null Hypothesis (H0)

The null hypothesis states that the mean **is equal to a hypothesized value/population mean**.

$$H0: \mu = 850$$

Alternative Hypothesis (H1)

The alternative hypothesis states that the mean **is smaller, greater, or different than the hypothesized value/population mean**.

One-Sided/One-Tailed

$$H1: \mu > 850$$

$$\mu < 850$$

Two-Sided/Two-Tailed

H1 : H1 :

$$\mu \neq 850$$

Test Statistic

A test statistic is **a random variable that is calculated from sample data** and **used in a hypothesis test**. We can use test statistics to determine whether to reject the null hypothesis (H0).

Since we run **t-test**, the test statistic is called **t-statistic.**

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

$\bar{X}$: sample's mean/average

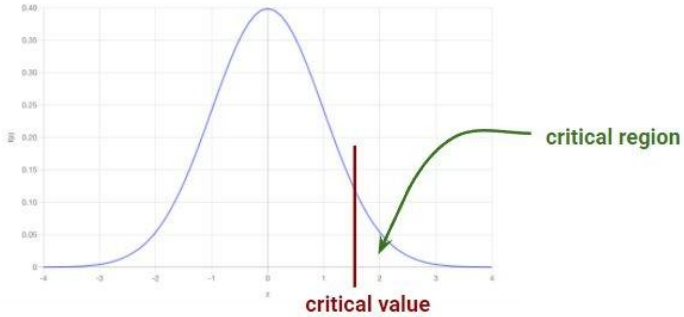
s : sample's standard deviation

μ : hypothesized value in H0 & H1

n : number of observations

Critical Value

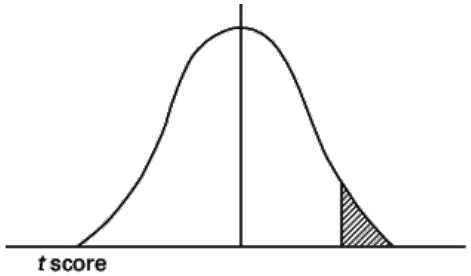
A critical value is **a point on the distribution of the test statistic under the null hypothesis that defines a set of values that call for rejecting the null hypothesis**.



Finding critical value for t-test:

1. Specify **Alpha (α)** for the test, usually: 5%
2. **Degrees of Freedom**, which is the number of observations in the sample (N) minus 1
3. Using **T-Table**, we can find the **corresponding T-Table value** (given alpha and degrees of freedom)

Critical Region



df \ p	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.683	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.160	2.650	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
24	1.318	1.711	2.064	2.492	2.797

H1 : $\mu > 850$

Critical Region: **t > 1.761**

Two-Sided/Two-Tailed

H1 : $\mu \neq 850$

Critical Region: **t < - 2.145 or t > 2.145**

Concluding T-Test

If the test statistic (T-statistic) that we have calculated is in the critical region:

We reject H0.
Conclusion: **Mean of the group is different/smaller/larger** than the hypothesized value or the population's mean.

If the test statistic (T-statistic) is outside the critical region:

We fail to reject H0.
Conclusion: **Mean of the group is equal to/the same as** the hypothesized value or the population's mean.

In Python

Performing T-Test:

```
import scipy.stats as stats

#perform one sample t-test

stats.ttest_1samp(a=data, popmean=mean)
```

(statistic=... pvalue=...)

Compare the P-value with alpha or compare the statistic with the critical region.

Real Case Example

Case Example

Company X wants to improve sales. Past sales data indicates that the average sale was $100 per transaction. After training the sales force, recent sales data (taken from a sample of 25 salesmen) indicates an average sale of $130, with a standard deviation of $15. Did the training work?
(Use alpha = 5%)

Step 1: Specify the hypotheses (H0 and H1)

H0: The average sale after training = $100

H1: The average sale after training > $100

Step 2: Specify the significance level (also called alpha or α) As shown in the question, we can use alpha = 0.05

Case Example

Company X wants to improve sales. Past sales data indicates that the average sale was $100 per transaction. After training the sales force, recent sales data (taken from a sample of 25 salesmen) indicates an average sale of $130, with a standard deviation of $15. Did the training work?
(Use alpha = 5%)

Step 3: Calculate T-statistic

$\bar{X}$: 130

μ : 100

s : 15

n : 25

$$t = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{130 - 100}{15/\sqrt{25}}$$

$$= \frac{30}{15/5} = 10$$

Case Example

Company X wants to improve sales. Past sales data indicates that the average sale was $100 per transaction. After training the sales force, recent sales data (taken from a sample of 25 salesmen) indicates an average sale of $130, with a standard deviation of $15. Did the training work?
(Use alpha = 5%)

Step 4: Find Critical Value (T-Table) and Critical

Region Degree of freedom = 24

Alpha = 0.05 (One-sided test)

Critical Value = 1.711
Critical Region =
t > 1.711

df \ p	0.1	0.05	0.025	0.01	0.005
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.683	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.160	2.650	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
24	1.318	1.711	2.064	2.492	2.797

Case Example

Company X wants to improve sales. Past sales data indicates that the average sale was $100 per transaction. After training the sales force, recent sales data (taken from a sample of 25 salesmen) indicates an average sale of $130, with a standard deviation of $15. Did the training work?
(Use alpha = 5%)

Step 5: Comparing and Make Conclusion

T-statistic = 10

Critical Region =
t > 1.711

Since t-statistic is in the critical region, then we reject H0.
Conclusion: The average sale after training > $100. The training was successful.

Assignment

Assignment 5

Kerjakan soal-soal pada file .ipynb yang terdapat pada assignment

Topic 9 & 10 - Assignment | Hands-On Python - Statistics pada Canvas, dimana terdapat kolom:

- Age and sex are self-explanatory
- BMI is body mass index
- BP is average blood pressure
- S1 through S6 are different blood measurements
- Y is the qualitative measure of disease progression over one year

kumpulkan link Google Colab dan beri nama notebook dengan format nama **Topik 9 10 - [Nama Lengkap]**

Available from	Until
Apr 16 at 03.00 PM	Apr 27 at 11.59PM

Thank you!

Any Questions?