Let’s Collaborate!

GOVINSIDER

SIGN UP FOR DAILY BRIEF

CONNECTED DIGITAL INCLUSIVE INNOVATION SECURITY SMART EVENTS ABOUT

Search GovInsider

This community is building Indonesia's data skills

Entrepreneur Fajar Jaman is working with civil servants and startups to boost data science skills in the country.

By Fahmi Ramadhan

2 DEC 2016

SMART GOV

Data Science Indonesia is a community working to advance data innovation across the nation. Pulse Lab Jakarta caught up with Fajar Jaman, one of the founders, to learn more about the network and its plans for the coming years.

What inspired you to establish Data Science Indonesia?

When I was working in Teradata, I was tasked to explore open data opportunities and to map the market in Indonesia. I was lucky to have my work align perfectly with my passion. I had a dream to initiate a community for data professionals who want to contribute to social projects.

I realised this dream in May 2015. We started with 30 members predominantly drawn from tech organisations, but now our community stands at 600 people including scientists, artists, civil servants, academics and students.

ABOUT US

Data Science Indonesia (DSI) was established in May 2015 to enable data ecosystem in Indonesia to create values for Indonesia society. This objective is done through three pillars of DSI: **educate** the people, **advocate** the data-driven culture, and establish strong **network** among data professionals

Currently, we are the biggest data community in Indonesia and Southeast Asia

With 9,000+ members

in 16 cities in Indonesia

We are nurture Indonesia's data ecosystem through our three pillars

*not only for Data Scientist but also all data enthusiast

In the past 4 years, we have done 50+ activities in collaboration with industry players and public institutions



Organizing trainings, seminars, and peer learning activities


Hosted the largest community-organized data conference in Indonesia

SOME OF OUR PAST EVENTS & PROJECTS

Data Science Weekend is the largest community-organized data conference in Indonesia

3-day Workshop, Seminars, and Exhibition

500 participants

15+ speakers

18 Partners

Partners include:

UMN

GO-JEK

NIOMETRICS

HOME CREDIT

Yellowfin

iykra

BukaLapak

aws

nodeflux

eFishery

Beritagar.id

CODEPOLITAN

BLOCK71

DailySocial.id

TECHINASIA

KATADATA

INFOKOMPUTER

Data Science Weekend Snapshots

Partners include:

UMN

GO-JEK

NIOMETRICS

HOME CREDIT

Yellowfin

iykra

BukaLapak

aws

nodeflux

eFishery

Beritagar

CODEPOLITAN

BLOCK71

DailySocial.id

TECHINASIA

KATADATA

INFOKOMPUTER

DSI Bootcamp was an intensive training in Data Science with comprehensive curriculum. This training was the first community organized data science bootcamp in the country

6 months of intensive training with capstone project

25 participants

10+ trainers

25 Invited organizations

Invited organizations:

SALE STOCK, adskom, Telkom Indonesia, eFishery, Lippo, BSG, nutrifood, niometrics, traveloka, The Asia Foundation, Australian AID, Hivos people unlimited, kudo, HYVE, Perludem, fabelio, RTI INTERNATIONAL, DEKA, KANTOR STAF PRESIDEN, WORLD RESOURCES INSTITUTE, SEPAM CONSULTANTS, AWRAGO

DSI was advocating Pemkot Bandung in solving traffic congestion problem in Bandung

Partnering with HIVOS Indonesia, DSI was using data innovation to solve congestion problem in Bandung

Executed by DSI members

Insights were used for policy-making process

DSI Meetup and DataTalk is a workshop or seminar event executed by DSI partnering with industry

During this event, DSI and industry will pick the theme together and choose who will be presenting

~100 participants per event

DSI Meetup is for DSI Member Only
DataTalk is open for public

Partnering with 15+ partners

Data Pods is official podcast of DSI that discussed data ecosystem in Indonesia

Currently we have 16 episodes with different guests discussing about all application of data, ranging from skin imaging to music recognition

~200 listeners

15 guests

16 episodes

https://linktr.ee/datascienceindo

@datascienceindo

Radio Streaming: Bicara Data (Biweekly setiap Rabu pk. 7pm)

Learning Aset Management DSI

Medium DSI

Datapods on Spotify

Forum DSI

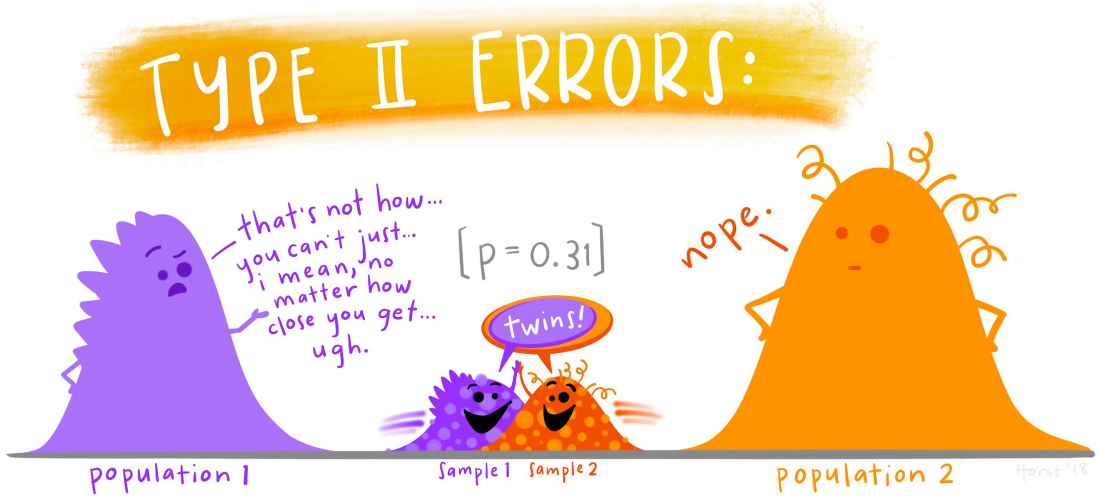
Daftar Member

@datascienceindo

@DataScienceIndo

Data Science Indonesia

Hypothesis Testing Using R



Artworks by @allison_horst

RLadies Jakarta 12
&
Data Science Indonesia

Hamidah Alatas

Today's Agenda

- **What is hypothesis testing?**
- **Hypothesis testing in R**

Hypothesis Testing

The management of a gym claims that its members lose an average of 10 kg or more within the first three months after joining the club.

A consumer agency that wanted to check this claim took a random sample of 36 members of this health club and found that they lost an average of 9.2 kg within the first month of membership with a standard deviation of 2.4 kg.

How can we sure the claim is correct?

Research hypothesis

Research hypothesis involves making a substantive, testable scientific claim.

Proposing a solution to answer your problem statement.

Example:

- Proper exercise method help you lose 10 kg within the first three months
- Listening to music reduces your ability to pay attention to other things.

Statistical hypothesis

Statistical hypotheses must be mathematically precise, and they must correspond to specific claims about the characteristics of the data generating mechanism (i.e., the “population”).

Example:

- Participants in my experiment lose weight (μ) = 10 kg

Hypothesis Formulation

- **Null Hypothesis (Ho)**: hypothesis that sample observations result purely from chance. Usually corresponds to the exact opposite of what we want to believe
- **Alternative hypothesis (Ha)**: hypothesis that sample observations are influenced by some non-random cause

Possible outcome:

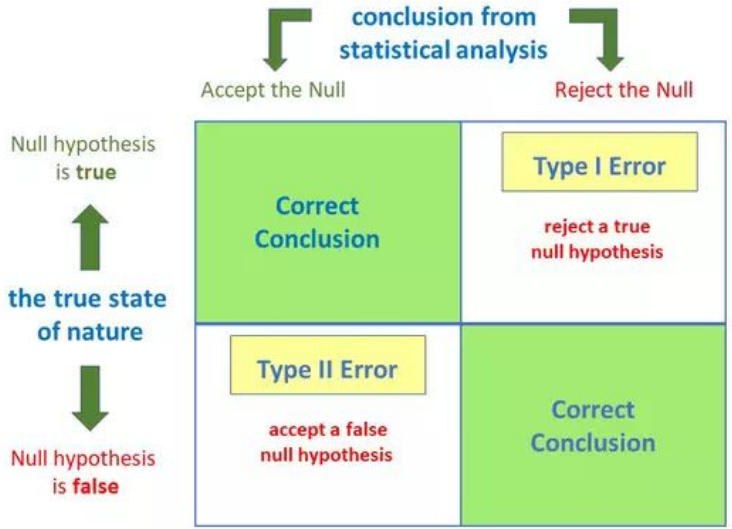
- Reject the Null Hypothesis
- Fail to reject the Null Hypothesis

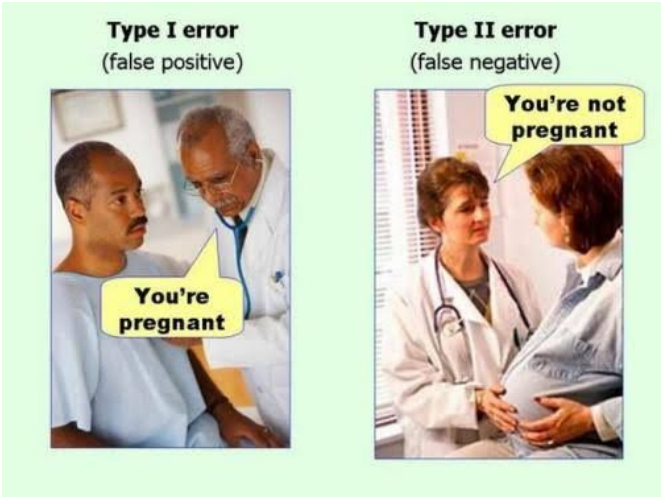
Example

- Ho: $\mu \neq 10$
- Ha: $\mu = 10$

Decision Making

In real life we always have to accept that there's a chance that we did the wrong thing. As a consequence, the goal behind statistical hypothesis testing is not to eliminate errors, but to minimise them.





Test Statistics

4 kg

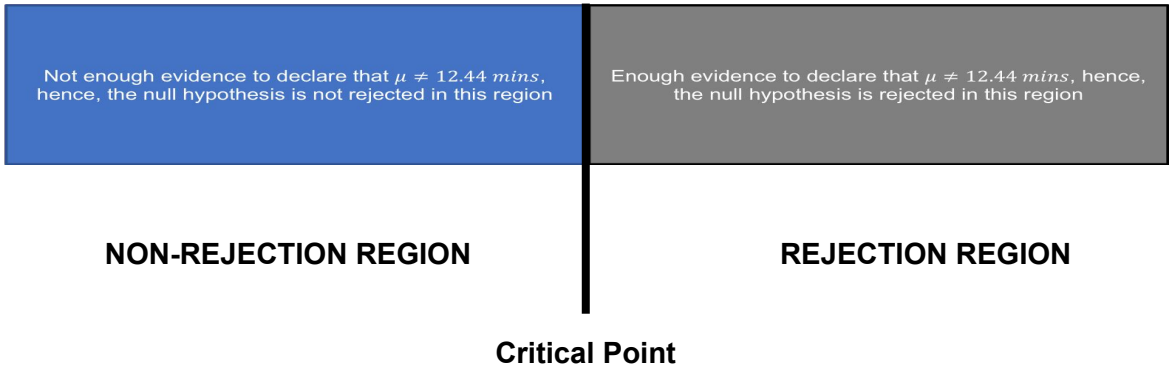
9 kg

15 kg

Test Statistics

Test statistic
After we sampled the data and define numerical outcome related to our hypotheses. It's used to decide whether to accept/reject Null Hypothesis.

Statistically significant
Where do we draw the line to make a decision based on the test statistic



We can use p-value to test the significance!

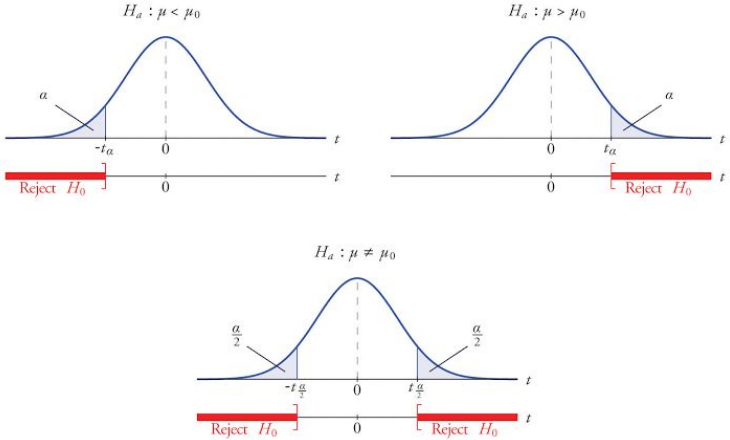
P-value can be defined as the probability obtaining a result that is “more extreme" than the one we observe if Null Hypothesis is true.

Or in simple word, p-values are numbers between 0 and 1 that quantify probability we do false positive error.

In other words, the closer p-value to 0, the more confident we are that we can reject the Null Hypothesis.

Tails of a Test

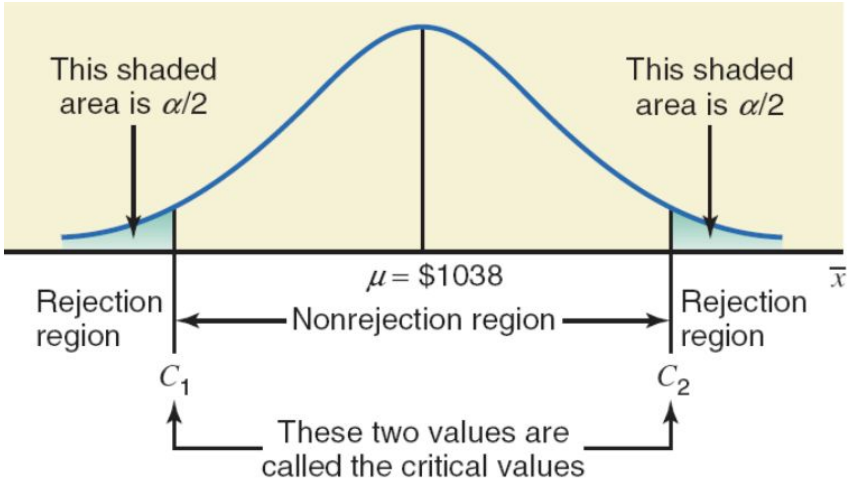
- **Two-tailed test** has rejection regions in both tails
- **Left-tailed test** has the rejection region in the left tail
- **Right-tailed test** has the rejection region in the right tail of the distribution curve.



Hypothesis Testing

- We reject the **Null Hypothesis** if we got our sample data in which the probability of committing a Type I error is lower than our standard (*significance level*, α)
- OR our test statistic is “more extreme” than critical values

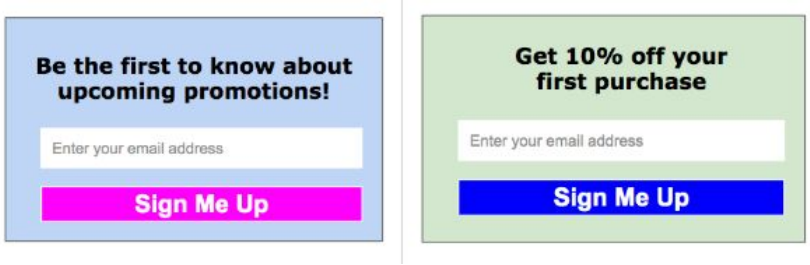
Null Hypothesis Distribution



Hypothesis testing popular use cases

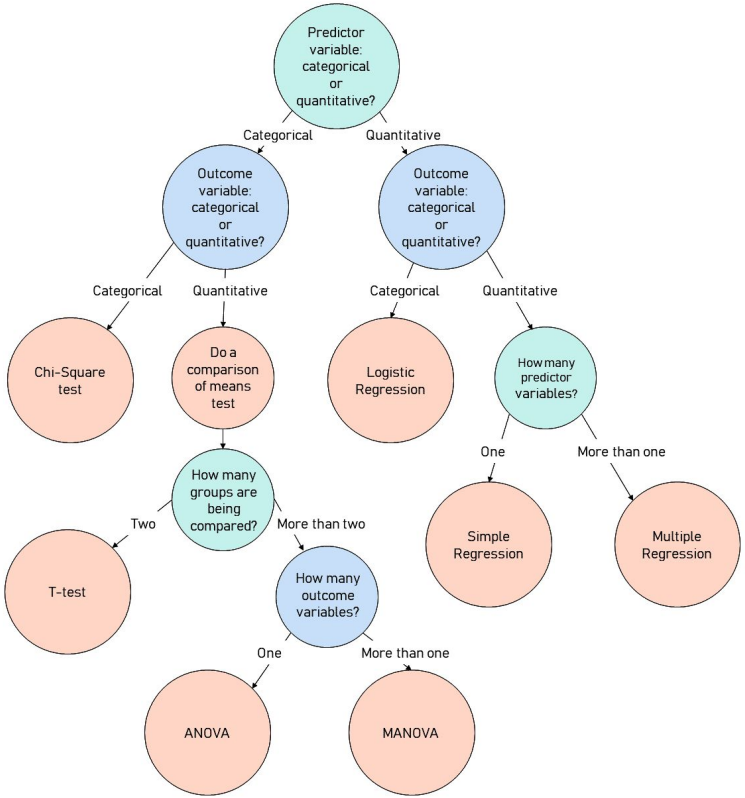
- **Verify assumption on sample statistics.**
  - Example: On average, members of gym A lose 10 kg

- **Testing differences in mean between two groups**
  - Example: A/B Testing between two banners in a webpage

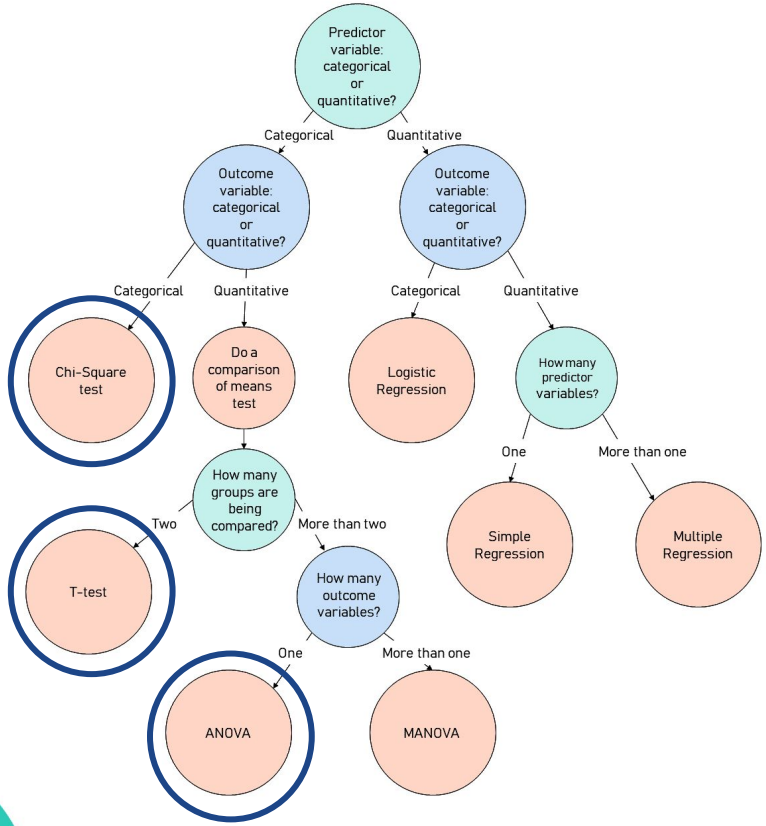


- Testing differences in mean before and after experiment
  - Example: Pre- and post- test after a training

Choosing a statistical test



Choosing a statistical test



Meet the test!

T-test is very common statistical test. It can be used for example to test if the means of two groups differ from each other statistically. You can also test if the mean of some variable is statistically different from predetermined value.

The t-test is done in R with the `t.test()` function. By default, the alternative hypothesis is "two.sided".

Assumption

- Normality. It is assumed that the data are normally distributed. Specifically, we assume that both groups are normally distributed.
- Independence. Firstly, we assume that the observations within each sample are independent of one another (exactly the same as for the one-sample test). However, we also assume that there are no cross-sample dependencies.

Meet the test! (2)

The **chi-square** test (also written as χ2-test) is well known statistical test. It can be used to test whether two variables are *independent* from each other (e.g. one variable is not affected by the presence of another). The test can also be used as goodness-of-fit test: with it we can check if the observed frequency distribution differs from a theoretical distribution.

In R you can do the chi-squared test with the function `chisq.test()`.

Assumption

- Expected frequencies are sufficiently large
- Data are independent of one another

Meet the test! (3)

The **ANOVA** test is well known statistical test to compare several means. We're talking about an analysis similar to the t-test but involving more than two groups

In R you can do the ANOVA test with the function `aov()`.

Assumption

- Normality. It is assumed that the data are normally distributed. Specifically, we assume that both groups are normally distributed.
- Independence. Firstly, we assume that the observations within each sample are independent of one another (exactly the same as for the one-sample test). However, we also assume that there are no cross-sample dependencies.
- Homogeneity of variance (also called "homoscedasticity"). The third assumption is that the population standard deviation is the same in both groups. You can test this assumption using the Levene test

Other Test

Continuous

- Paired t-test. t-test but data are dependent.
- Wilcox. t-test but data not normally distributed.
- Kruskal-Wallis. ANOVA but data not normally distributed

Categorical

- Fischer test. Chi-squared test but sample are small.
- McNemar. Chi-squared test but data are dependent.

Steps

- Define Ho and Ha
- Check assumptions
- Choose test-statistics
- Inspect p-value
- Make the decision

Thank You!

@hamidahamidah

hamidah.alatas@gmail.com

Hamidah Alatas

Source

- Learning Statistics with R by Danielle Navarro
- Lecture 7 - Hypothesis testing in R by Prof. Alexandra Chouldechova