



# Advanced Statistic



# Table of Content

## What will We Learn Today?

1. Sampling
2. Hypothesis Testing
3. AB Testing
4. Platform of AB Testing
5. Practice





# Profile




## Professional

- Senior Data Analyst – Kompas (2021 – Present)
- Data Scientist – Rukita (2020 – 2021)
- Research Assistant Analyst – Ensterna (2017 – 2019)

## Educational Background

- Nuclear Engineering – Universitas Gadjah Mada

## Connect with me

-  <https://dataimpact.medium.com/>
-  <https://www.linkedin.com/in/ariprabowo/>
-  <https://github.com/densaiko>



**Ari Sulistiyo Prabowo**





# Sampling

*Sampling data mengacu pada metode statistik untuk memilih pengamatan dengan tujuan memperkirakan parameter populasi*

# Sampling

**Jika kita ingin mengetahui perilaku customer, kita seringkali tidak memiliki akses untuk kemungkinan yang ada untuk seluruh datanya. Mengapa?**

01

Mengumpulkan seluruh data akan sangat sulit, mahal dan memakan waktu yang banyak

02

Observasi lanjutan dapat dilakukan jika sampling belum terpenuhi

03

Pengembangan data di kemudian hari untuk analisis/penelitian lain





# How to sample

Dalam melakukan sampling, ada beberapa aspek yang perlu dipertimbangkan sebelum mengumpulkan data:

- **Tujuan sampel:** Bagian dari populasi yang ingin anda perkirakan
- **Population:** Ruang lingkup dari mana pengamatan anda dimulai
- **Kriteria Seleksi:** Metodologi yang digunakan untuk mengambil spesifik informasi dari observasi
- **Ukuran sampel:** banyaknya pengamatan yang akan dijadikan sampel







# How to sample

Pengambilan sampel statistik adalah bidang studi yang luas, tapi dalam pembelajaran “applied machine learning”, terdapat tiga jenis teknik sampling:

- **Simple Random Sampling:** Sampel yang diambil dengan probabilitas seragam dari populasi
- **Systematic Sampling:** Sampel yang diambil menggunakan pola yang ditentukan sebelumnya dengan bantuan interval
- **Stratified Sampling:** Sampel yang diambil dengan kategori yang ditentukan



# Hypothesis Testing

*Digunakan untuk melakukan praduga/prediksi/hipotesa/dugaan dari populasi dan sampel data yang ada*







# Tujuan Hypothesis Testing

Tes statistik hipotesis didasarkan pada sebuah statement yang disebut **null hypothesis** yang artinya anda menebak sesuatu pada data

- **Untuk menentukan** apakah null hypothesis kemungkinan benar dari perkiraan awal
- **Untuk menganalisis** apakah anda memerlukan bukti untuk naik ke level berikutnya yaitu AB testing

## Before going further!

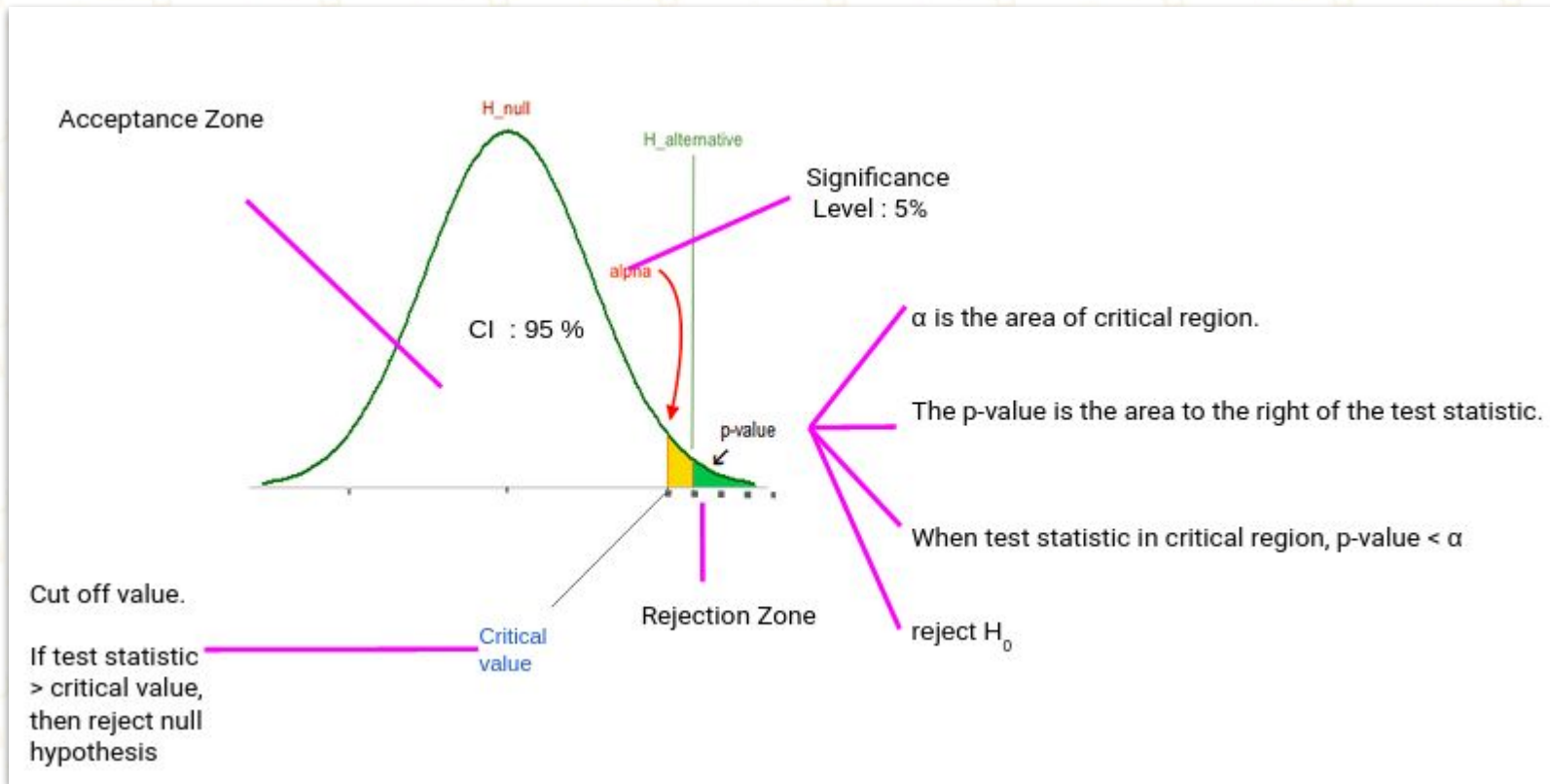


# Anda harus memahami ini!

- **Ho (null hypothesis)** apakah dugaan awal dapat dibenarkan dari sampel data
- **Ha (alternative hypothesis)** apakah anda membutuhkan bukti untuk menolak dugaan awal yang dimana dapat ke langkah berikutnya yaitu AB testing
- **Tingkat kepercayaan (confidence level)** adalah metrik yang anda tentukan untuk mempercayai hasil dari dugaan anda
- **$\alpha$  (significant level)** (1-confidence level) adalah potongan nilai antara zona menerima hipotesis dan menolak hipotesis
- **p-value (critical value)** adalah nilai dari tes hipotesis guna untuk menerima atau menolak null hypothesis



# You need to understand!!







# Hypothesis testing (Z-test & T-test)

## Z-test

### Properties:

- Diketahui varians dari populasi
- Jika tidak ada varians dari populasi, ukuran sampel harus melebihi 30 data

sample mean

Population mean

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

Population standard  
deviation

Sample size



# Case Z-test

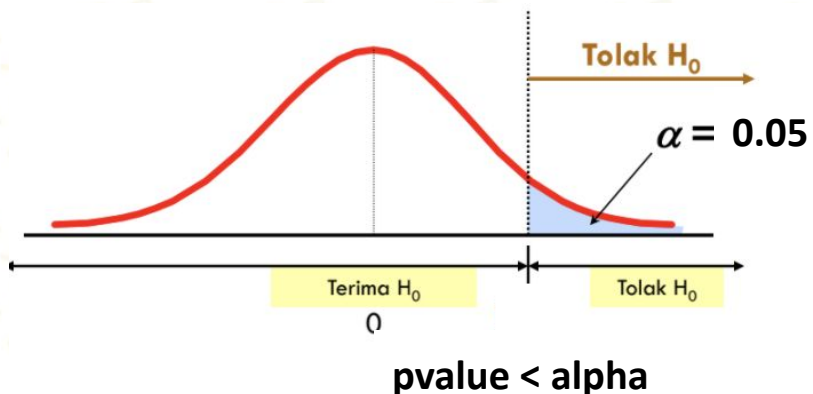
Transformer memiliki rata-rata CO2 sebesar 1186, pada saat bekerja transformer apakah CO2 yang dihasilkan lebih tinggi dari rata-rata CO2? Sampel yang diambil adalah 35 data,

Hipotesis:

- $H_0$ : rata-rata  $> 1186$ , rata-ratanya lebih besar dari 1186
- $H_1$ : rata-rata  $< 1186$ , rata-ratanya tidak lebih besar dari 1186

Jawabannya:

Misalkan  $\alpha = 0.05$  yang digunakan untuk uji hipotesis ini dan  $n = 35$ , maka area-nya sebagai berikut





# Hypothesis testing (Z-test & T-test)

## T-test

### Properties:

- Tidak diketahui varians populasi
- Jumlah sampel data kecil,  $n \leq 30$

sample mean

Population mean

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

Sample standard  
deviation

Sample size





# Case t-test

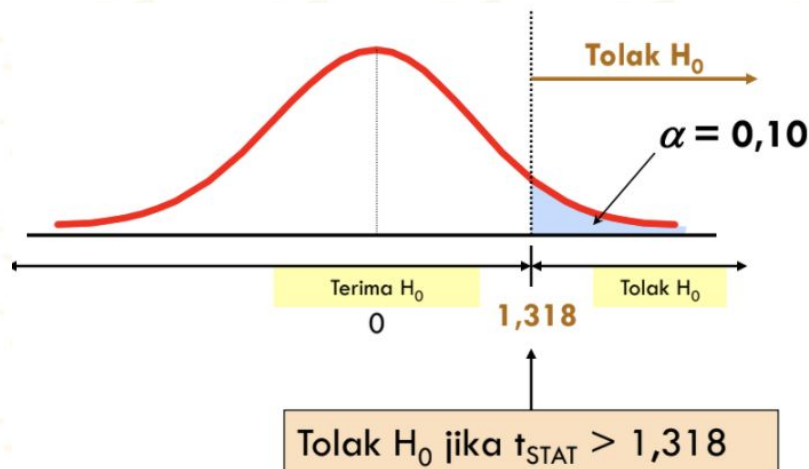
Seorang manajer penyedia layanan telepon selular berpendapat bahwa telah terjadi peningkatan tagihan telepon pelanggan, sehingga **rata-ratanya menjadi lebih dari \$52 per bulan**. Perusahaan ingin menguji pernyataan ini. Terdapat 25 sampel. (Diasumsikan populasi berdistribusi normal)

Hipotesis:

- $H_0$ : rata-rata  $\leq 52$ , rata-ratanya tidak lebih dari \$52 per bulan
- $H_1$ : rata-rata  $> 52$ , rata-ratanya lebih dari \$52 per bulan

Jawabannya:

Misalkan  $\alpha = 0.1$  yang digunakan untuk uji hipotesis ini dan  $n = 25$ , maka area-nya sebagai berikut





# Keputusan Hypothesis Testing



**Mba Catherine**

Jika nilai  $p\text{-value} < \alpha$  maka tolak  $H_0$ ,  
jika sebaliknya, terima  $H_0$

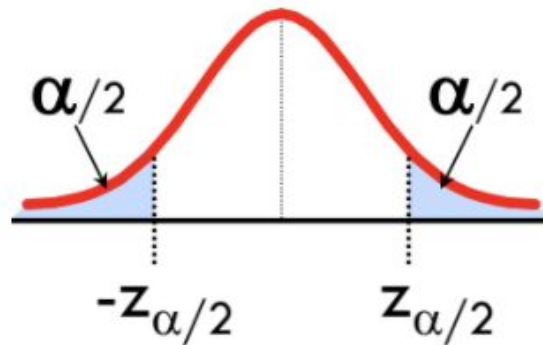


# Hypothesis Testing - Two Sample

Two-tail test:

$$H_0: \pi_1 - \pi_2 = 0$$

$$H_1: \pi_1 - \pi_2 \neq 0$$



Tolak  $H_0$  jika  $Z_{STAT} < -Z_{\alpha/2}$   
atau  $Z_{STAT} > Z_{\alpha/2}$





# Case two sample

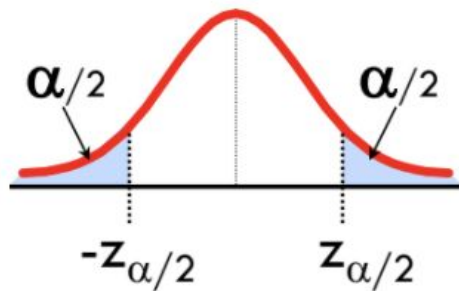
Transformer memiliki karakter metana yang dihasilkan, ketika transformer bekerja dan tidak bekerja apakah rata-rata metana yang dihasilkan sama?

Hipotesis:

- $H_0$ : rata-rata metana saat bekerja = rata-rata metana saat tidak bekerja
- $H_1$ : rata-rata metana saat bekerja  $\neq$  rata-rata metana saat tidak bekerja

Jawabannya:

Misalkan  $\alpha = 0.05$  yang digunakan untuk uji hipotesis ini dan  $n = 30$ , maka area-nya sebagai berikut



Tolak  $H_0$  jika  $Z_{STAT} < -Z_{\alpha/2}$   
atau  $Z_{STAT} > Z_{\alpha/2}$



# Hypothesis Testing - Chi Square

Hipotesis statistik menggunakan chi square bertujuan untuk melakukan prediksi secara statistik (tes hipotesa) dengan melihat ada atau tidaknya hubungan diantara beberapa variabel

Contoh: Pada transformer listrik, terdapat beberapa senyawa yang dihasilkan seperti CO, CO<sub>2</sub>, Hydrogen dan Metana. Pertanyaannya adalah, apakah kegiatan transformer listrik memiliki hubungan dengan beberapa senyawa tersebut?

Hipotesis:

- H<sub>0</sub>: variabel tersebut berhubungan satu sama lainnya (dependent)
- H<sub>1</sub>: Variabel tersebut tidak berhubungan satu sama lainnya (independent)

**Check on the Notebook**



# AB Testing

*Menguji varian pengamatan anda dan melihat bagaimana kinerjanya terhadap tujuan yang anda tentukan*



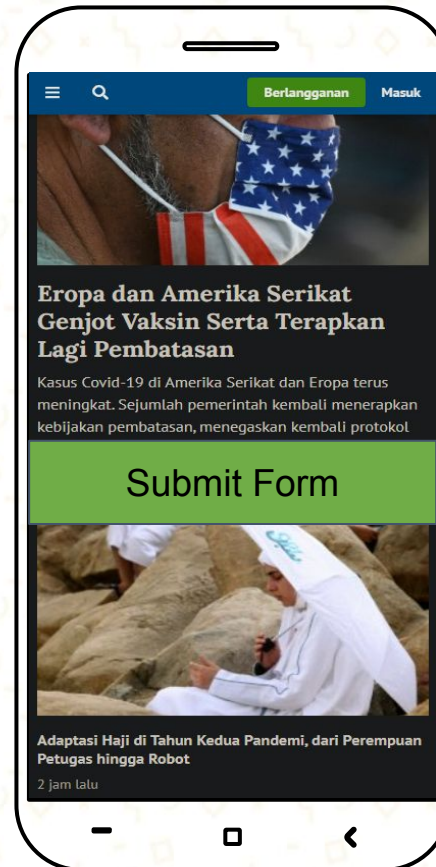


# Example of AB Testing

A



B



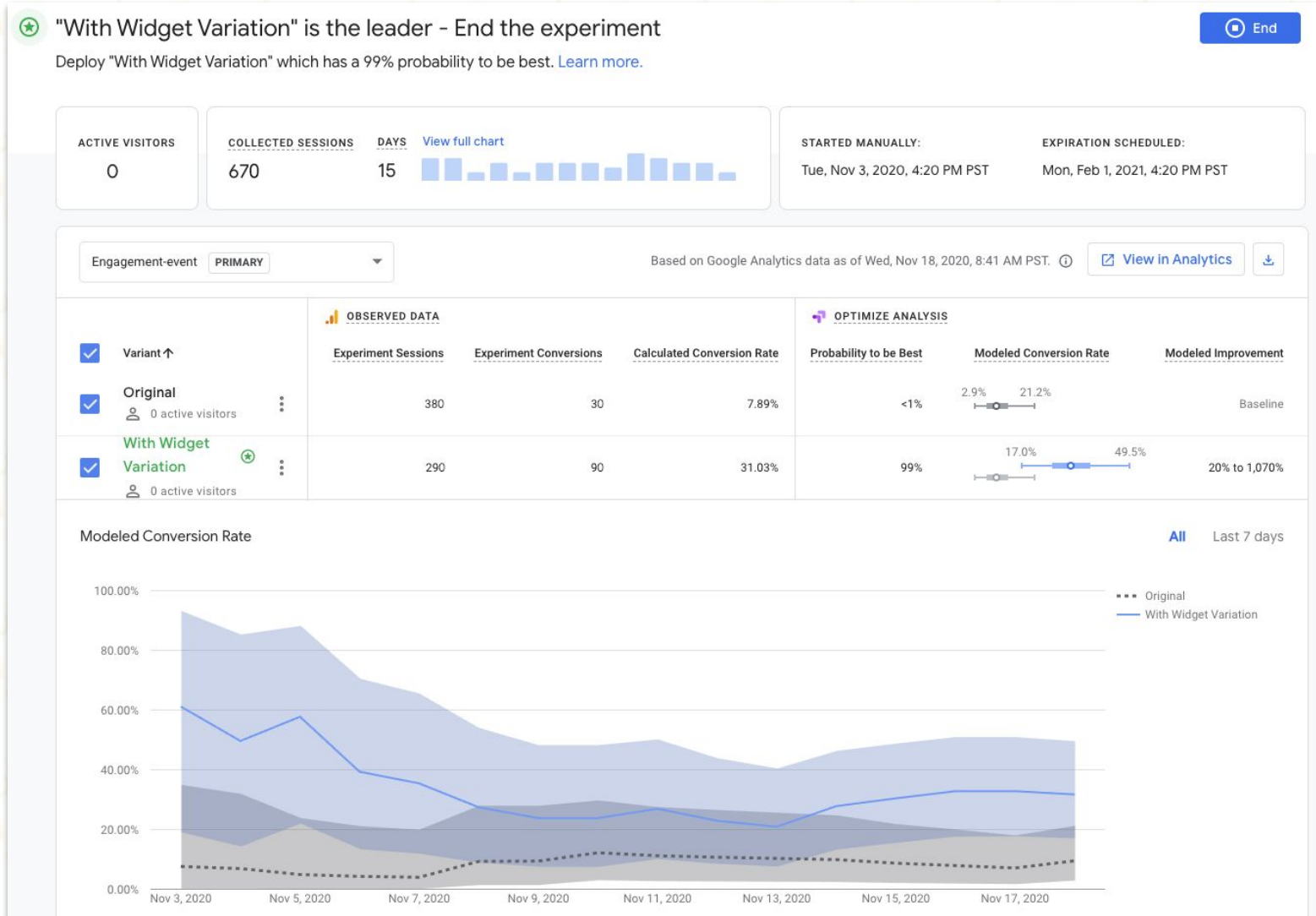


# Step-by-step AB Testing in Website

- **Apa tujuan anda?** mendefinisikan tujuan dari eksperimen anda. Contohnya adalah jumlah klik, jumlah submit, user dan pageview
- **Apa hipotesis anda?** dengan menambahkan form submit antara kedua artikel dapat meningkatkan jumlah user dalam subscribe
- **UI/UX team and front end engineer** meminta mereka dalam mendesain fitur A dan B dalam website anda
- **Setting platform AB testing** sebagai seorang data scientist/data analyst perlu paham untuk memasang tracker untuk mengambil data
- **Pada pengaturan platform AB testing** anda dapat melakukan setting 50% - 50% (usually), dan setingan lainnya terkait data apa yang ingin diperoleh



# Platforms of AB Testing





# Homework

Dataset: [https://raw.githubusercontent.com/densaiko/data\\_science\\_learning/main/dataset/BankChurners.csv](https://raw.githubusercontent.com/densaiko/data_science_learning/main/dataset/BankChurners.csv)

Dalam dataset the Bank Churn, **Credit limit** merupakan salah satu variable untuk menentukan apakah pelanggan akan churn/attrition (berhenti) atau masih tetap menggunakan produk. Lakukan hipotesis **apakah rata-rata credit limit laki-laki dan perempuan sama?**

- Ambil 50 samples
- Tingkat kepercayaan (confidence interval) 95%
- Jenis tes mana yang anda pilih? z-test atau t-test, mengapa?
- Apakah anda menggunakan one sample atau two sample?



**Thank  
YOU**

