Practical Statistics

December 26, 2023

Coursebook: Practical Statistics

- Bagian 2 Audit Analytics untuk Bank Rakyat Indonesia
- Durasi: 7 Jam
- *Last Updated*: December 2023

- *Coursebook* ini disusun dan dikurasi oleh tim produk dan instruktur dari Algoritma Data Science School

1 Background

Coursebook ini merupakan bagian dari **BRI Audit Analytics** yang disiapkan oleh Algoritma. *Coursebook* ini ditujukan hanya untuk khalayak terbatas, yaitu individu dan organisasi yang menerima *coursebook* ini langsung dari organisasi pelatihan. Tidak boleh direproduksi, didistribusikan, diterjemahkan, atau diadaptasi dalam bentuk apapun di luar individu dan organisasi ini tanpa izin.

Algoritma adalah pusat pendidikan *data science* yang berbasis di Jakarta. Kami menyelenggarakan *workshop* dan program pelatihan untuk membantu para profesional dan mahasiswa dalam menguasai berbagai sub-bidang *data science* yaitu: *data visualization, machine learning,* statistik, dan lain sebagainya.

2 Training Objectives

Descriptive Statistics

- Understanding 5 number summary
- Central tendency measure
- Measure of spread
- Variable relationship
- Z Score and Central Limit Theorem

Inferential Statistics

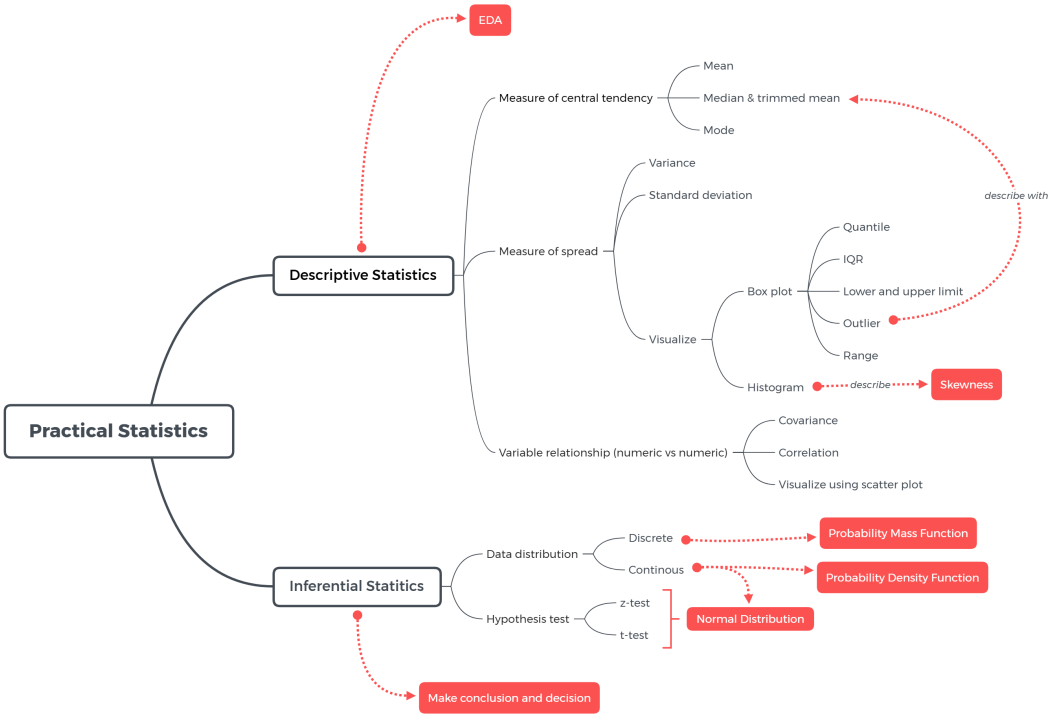
- Probability Density Function
- Data distributions
- Hypothesis test
- Error and confidence intervals

3 Practical Statistics

Practical Statistics merupakan salah satu bagian penting dalam pengolahan data sehingga mendukung *busines case* yang ingin diangkat. Practical Statistics berisi kaidah statistika yang banyak diterapkan dalam praktik data science agar dapat **memahami dan mengolah data dengan tepat**.

Secara umum, Practical Statistics terbagi 2, masing-masing membantu kita dalam hal tertentu:

- **Descriptive Statistics**: meringkas informasi dalam data agar terambil insight secara cepat. Nilai-nilai yang didapatkan merupakan rangkuman dari data, tujuannya untuk menggambarkan keadaan data secara umum.
- **Inferential Statistics**: menyimpulkan sesuatu tentang kondisi di lapangan, berdasarkan data yang kita punya (sample -> population).



Untuk lebih memahami *practical statistics*, mari kita melakukan analisis menggunakan data asli.

[1]:
```
import pandas as pd
import numpy as np #perhitungan statistik
import matplotlib.pyplot as plt # untuk visualisasi
import math # perhitungan statistik
from scipy import stats #untuk perhitungan statistik
import seaborn as sns # untuk visualisasi
```

4 Study Case: Credit Card Balance Analysis

1. Business Question

Credit Card Balance Analysis, atau Analisis Saldo Kartu Kredit, dilakukan sebagai bagian dari analisis debitur dalam sebuah perusahaan kartu kredit. Hasil analisis dapat menentukan debitur mana yang memiliki risiko pembayaran kredit yang tinggi, atau bagaimana behavior debitur. Selain itu, menggabungkan data saldo kredit dengan informasi seperti limit kredit dapat membantu menghitung pemanfaatan kredit kartu, informasi yang berpengaruh pada Rating kredit seorang pemegang kartu.

Asumsi data:

- Balance dihitung sebagai jumlah semua transaksi selama periode penagihan/billing cycle. Sebagai contoh, jika seorang pemegang kartu mengeluarkan $400, $500, dan $600 dalam 3 bulan, maka saldo rata-rata akan dicatat sebagai $500.

Kita sebagai tim data diminta untuk **menganalisa performa Credit Card Balance** nasabah. Data tersimpan dalam folder data_input dengan nama file `credit_card.csv`, gunakan `stringAsFactors = T` supaya kolom bernilai string berubah langsung menjadi tipe factor.

2. Read Data

[2]:
```
cc = pd.read_csv('data_input/CC.csv')
cc.head()
```

[2]:

	Income	Limit	Rating	Cards	Age	Education	Gender	Student	Married	\
0	14.891	3606	283	2	34	11	Male	No	Yes	
1	106.025	6645	483	3	82	15	Female	Yes	Yes	
2	104.593	7075	514	4	71	11	Male	No	No	
3	148.924	9504	681	3	36	11	Female	No	No	
4	55.882	4897	357	2	68	16	Male	No	Yes	

	Ethnicity	Balance
0	Caucasian	333
1	Asian	903
2	Asian	580
3	Asian	964
4	Caucasian	331

Deskripsi Kolom

- `Income`: Besaran gaji nasabah per tahun (dalam $10000)
- `Limit` : Besaran kredit limit
- `Rating` : Skor yang diberikan kepada individu berdasarkan kelayakan kreditnya. Semakin besar maka semakin baik
- `Cards` : Jumlah banyaknya kartu kredit yang dimiliki oleh nasabah
- `Age` : Usia nasabah
- `Education` : Level/lamanya pendidikan yang ditempuh oleh nasabah
- `Gender`: Jenis kelamin nasabah
  - Male

  - Female
- `Student` : Apakah nasabah seorang pelajar atau bukan
  - Yes → Pelajar
  - No → Bukan pelajar
- `Married`: Status pernikahan
  - Yes → Sudah menikah
  - No → Belum menikah
- `Ethnicity`: Etnis nasabah
  - African American
  - Asian
  - Caucasian
- `Balance`: Rata-rata jumlah saldo kartu kredit

5 Descriptive Statistics

Descriptive Statistics membantu kita **menggambarkan karakteristik** dari data, sehingga berguna dalam proses **Exploratory Data Analysis (EDA)**. Terdapat 3 hal pada descriptive statistics:

- Ukuran pemusatan data (Measure of Central Tendency)
- Ukuran penyebaran data (Measure of Spread)
- Hubungan antar data (Variable Relationship)

5.1 Measure of Central Tendency

Ukuran pemusatan data adalah **suatu nilai yang cukup untuk mewakili seluruh nilai pada data**.

5.1.1 Mean

Cara paling umum untuk membuat perkiraan nilai tunggal dari data yang banyak adalah dengan merata-ratakannya.

- Formula:

$$\frac{\sum x_i}{n}$$

- Fungsi pada Python: `mean()`

Contoh:

Berapa rata-rata `Rating` atau skor yang diberikan kepada nasabah berdasarkan kelayakan kreditnya?

[14]:
```
# code here


# end of the code
```

- Sifat nilai mean: **sensitif terhadap outlier**

Outlier adalah nilai ekstrim yang jauh dari observasi lainnya. Kurang tepat apabila menggunakan nilai mean yang diketahui ada data outliernya.

Contoh lain:

Ada sebuah Kantor Cabang BRI di daerah Bekasi yang merekap jumlah pengunjung per bulan.

Dengan nilai mean:

[7]:
```
# data pengunjung
pengunjung = pd.Series([55, 50, 40, 70, 60, 45, 35, 35, 60, 1000, 250, 70])
```

[13]:
```
# rata-rata pengunjung


# end of the code
```

Apakah nilai mean di atas dapat diandalkan? _____

Nilai mean tidak dapat diandalkan karena terdapat outlier

Masalah ini dapat diatasi oleh nilai **median**.

5.1.2 Median

Median atau nilai tengah diperoleh dengan mengurutkan data terlebih dahulu kemudian mencari nilai tengah dari data.

- Baik untuk data yang memiliki **outlier** atau berdistribusi **skewed** (condong kiri/kanan)
- Fungsi pada Python: `median()`

Mari hitung ulang nilai pusat dari `pengunjung` menggunakan median:

[9]:
```
# median


# end of the code
```

[9]:
```
147.5
```

[10]:
```
# bandingkan dengan mean


# end of the code
```

[10]:
```
147.5
```

Untuk nilai desimal, apabila tidak sesuai dengan konteks bisnis (Pengunjung), chaining hasil dengan fungsi rounding `round()`

[11]:
```
pengunjung.mean().round()
```

[11]:
```
148.0
```

5.1.3 Modus (Mode)

Modus berguna untuk mencari nilai yang paling sering muncul (frekuensi tertinggi).

- Modus digunakan untuk data kategorik
- Fungsi pada Python: `mode()`

Contoh:

Berasal dari `Ethnicity` mana nasabah di Bank tersebut paling banyak berasal?

[12]:
```
cc['Ethnicity'].mode()
```

[12]:
```
0    Caucasian
Name: Ethnicity, dtype: object
```

Modus untuk `Ethnicity` adalah _____

5.1.4 Knowledge Check

Dari pernyataan berikut, jawablah benar atau salah. Apabila salah tuliskan pernyataan yang benar.

1. Median adalah pusat data yang hanya melibatkan sebagian data dalam perhitungannya.

☐ Benar
☐ Salah

2. Mean adalah pusat data yang sensitif terhadap outlier.

☐ Benar
☐ Salah

3. Nilai pusat data yang cocok untuk tipe data kategorik adalah modus.

☐ Benar
☐ Salah

5.2 Measure of Spread

Ukuran penyebaran data mewakili seberapa menyebar atau beragam data kita.

5.2.1 Variance

Variance menggambarkan seberapa beragam suatu data numerik tunggal menyebar dari pusat datanya

- Formula variance:

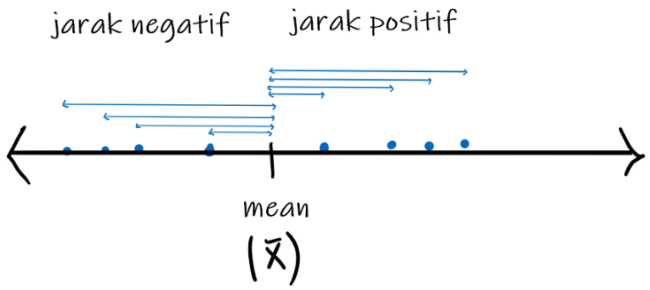
$$var = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

- Variance tidak dapat diinterpretasikan karena satuannya dalam kuadrat.
- Fungsi di Python: `var()`

$$var = \frac{\sum(X_i - \bar{X})^2}{n-1}$$

selisih jarak, kuadrat, rata²

$$sd = \sqrt{var}$$



Contoh:

Bank BRI sedang dalam rencana membuka kantor cabang baru. Bank BRI menyeleksi daerah mana yang cocok untuk cabang baru mereka. Mereka mengumpulkan informasi harga sewa bangunan di daerah A dan B sebagai berikut:

[15]:
```
# harga dalam satuan juta
harga_A = pd.Series([400,410,420,400,410,420,400,410,420,400,410,420,400])
harga_B = pd.Series([130,430,650,540,460,320,380,550,650,470,330,140,270])
```

Bandingkan rata-rata harga bangunan kedua daerah:

[19]:
```
# code here


# end of the code
```

Mari bandingkan dari sisi lain, yaitu tingkat keberagaman data (variance). Daerah mana yang harganya lebih bervariasi?

[18]:
```
# code here


# end of the code
```

Daerah manakah yang lebih baik untuk dijadikan area perkantoran?

...

Karakteristik Variance

- Skala variance dari 0 sampai tak hingga. Semakin besar nilainya maka artinya semakin menyebar dari pusat datanya (mean).
- Variance memiliki satuan kuadrat, sehingga tidak dapat langsung diinterpretasikan. Biasanya digunakan untuk membandingkan dengan nilai var lain dengan satuan yang sama.
- **Nilai variansi sangat bergantung dengan skala data**. Hati-hati apabila membandingkan antar variabel yang berbeda skala.

5.2.2 Standard Deviation

Standard deviation menggambarkan **seberapa jauh simpangan nilai yang dianggap umum, dihitung dari titik pusat (mean) nya.** Kita dapat menentukan apakah suatu nilai dikatakan menyimpang dari rata-rata namun masih dikatakan umum, atau sudah tidak umum.

Karena dihitung dengan **mengakarkan variance**, satuannya sudah sesuai dengan data asli dan bisa diinterpretasikan.

- Formula standar deviasi:

$$sd = \sqrt{var}$$

- Fungsi di Python: `std()`

[20]:
```
# standar deviasi harga_A & harga_B


# end of code
```

[21]:
```
# tinjau nilai mean harga_A & harga_B


# end of the code
```

Interpretasi nilai normal/wajar : mean +- sd (karena satuan mean dan sd sama, yaitu jutaan rupiah)

- Harga sewa pada daerah A umumnya jatuh pada interval _____
- Harga sewa pada daerah B umumnya jatuh pada interval _____

Business question

Apabila kita ditawarkan suatu bangunan di daerah B dengan harga 800, apakah harga tersebut masih wajar? Apakah sebaiknya kita membeli bangunan tersebut? Hubungkan dengan nilai mean dan standar deviasi yang diperoleh.

Hitung range "harga normal" daerah B:

[ ]:
```
# code here


# end of the code
```

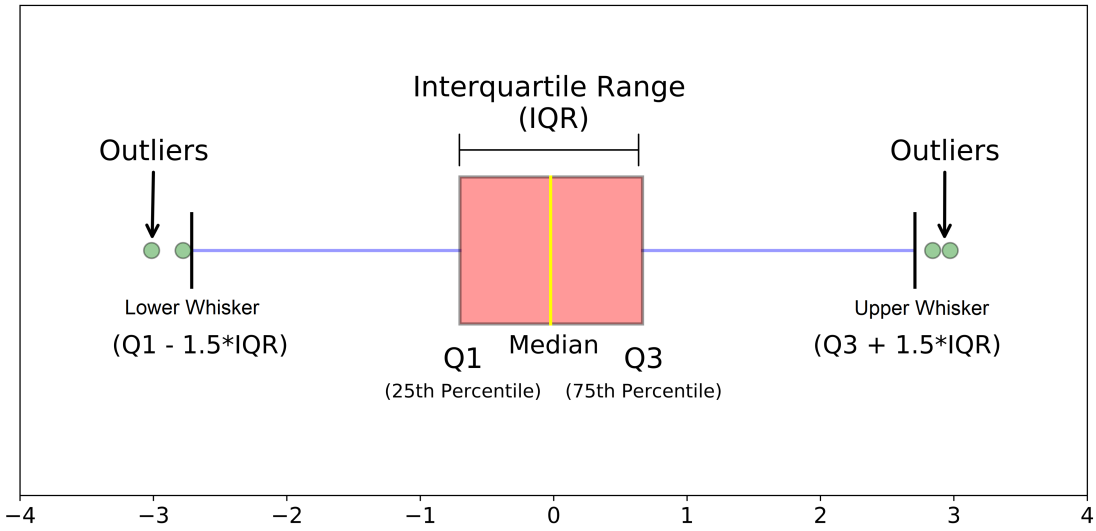
...

5.2.3 Range using `boxplot()`

Distribusi data numerik pada umumnya divisualisasikan dengan `boxplot()`, yang meliputi komponen:

- Box: menggambarkan Q1, Q2 (median), dan Q3
  - Kuartil 1 (Q1): nilai ke 25%
  - Kuartil 2 (Q2 atau median): nilai ke 50% (nilai tengah)
  - Kuartil 3 (Q3): nilai ke 75%

  - Interquartile Range (IQR): selisih antara Q3 dan Q1
- Whisker: pagar bawah dan atas (PENTING: hati-hati, nilai ini bukan nilai minimum dan maksimum data)
- Data outliers: nilai ekstrim data yang berada di luar pagar bawah dan atas



Beberapa hal yang harus diperhatikan dalam boxplot:

- Banyaknya data dari Q1 ke nilai minimum (bukan pagar bawah) adalah 25%
- Banyaknya data dari Q1 ke Q2 adalah 25%
- Banyaknya data dari Q2 ke Q3 adalah 25%
- Banyaknya data dari Q3 ke nilai maksimum (bukan pagar atas) adalah 25%

Insight yang dapat diperoleh dari boxplot:

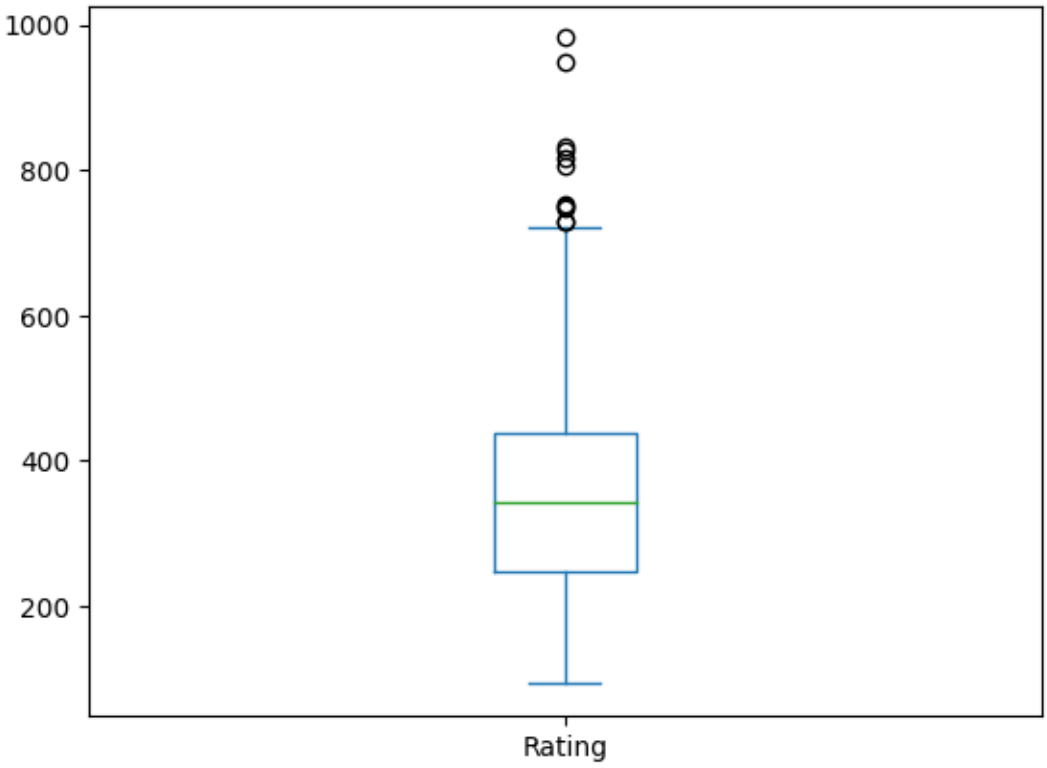
1. Pusat data dengan median (Q2)
2. Sebaran data dengan IQR (lebar kotak)
3. Outlier, nilai ekstrim pada data
4. Bentuk distribusi data:

- box yang berada ditengah = **distribusi normal**
- box yang mendekati batas bawah = **distribusi skewed kanan**
- box yang mendekati batas atas = **distribusi skewed kiri**

Contoh:

Visualisasikan sebaran data `Rating` dari data `cc`! Analisis informasi yang didapatkan.

[23]:
```
cc['Rating'].plot.box();
```



- Apakah data memiliki outlier?

  ...

- Central tendency (mean, median, modus) mana yang cocok dipakai untuk data ini?

  ...

- Bagaimana bentuk distribusi data?

  ...

5.3 Variable Relationship

Karena pada data kita punya banyak kolom atau variabel, kita juga ingin tahu hubungan antar variabel dalam data kita.

Ukuran yang digunakan untuk melihat **hubungan linear** antara dua variabel numerik.

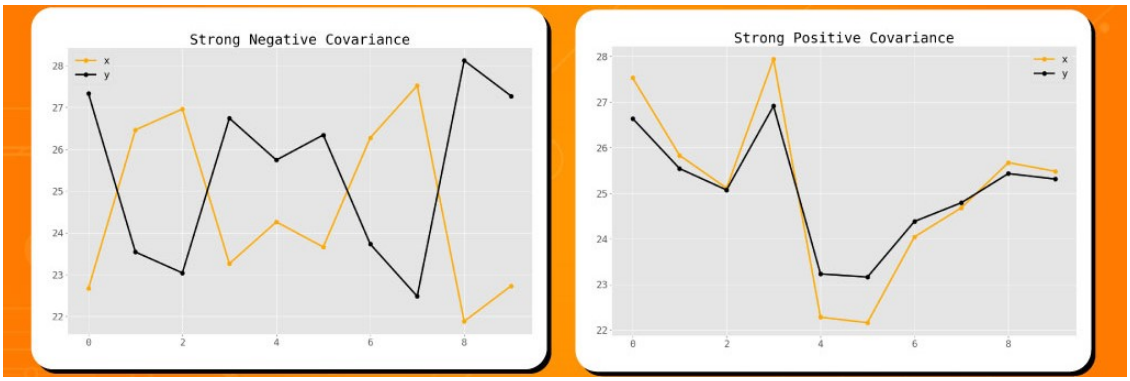
5.3.1 Covariance

Covariance menunjukkan bagaimana variansi 2 data (variable yang berbeda) bergerak bersamaan

- Formula Covariance:

$$Cov(X, Y) = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \mu_X)(Y_i - \mu_Y)$$

- Fungsi di Python: `cov()`



- Nilai covariance positif mengindikasikan pergerakan nilai yang searah / berbanding lurus.
- Nilai covariance negatif mengindikasikan pergerakan nilai yang berbalik arah.

Contoh:

Hitunglah covariance antara `Income` dengan `Rating` pada data `cc` . Bagaimana hubungannya?

```
[ ]: # code here


     # end of code
```

Interpretasi nilai: _____

Kelemahan: Seperti variance, covariance tidak memiliki batasan nilai untuk mengukur kekuatan hubungan antar dua variabel (-inf s.d inf), sehingga kita hanya bisa mengetahui apakah hubungannya positif atau negatif. Oleh karena itu, hadir **correlation**.

5.3.2 Correlation

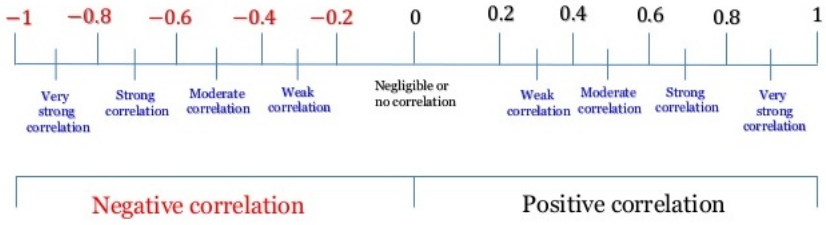
Correlation memampatkan nilai covariance dari -inf s.d inf menjadi **-1 s.d 1** sehingga bisa diukur kekuatan hubungan antar data (variable).

- Formula Correlation:

$$Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$$

* Fungsi di Python: `corr()`

- Nilai korelasi mengindikasikan kekuatan hubungan antara dua variable numerik sebagai berikut:



Bila korelasi dua variable numerik mendekati: - -1 artinya korelasi negatif kuat - 0 artinya tidak berkorelasi - 1 artinya korelasi positif kuat

Contoh:

Adakah korelasi antara `Income` dengan `Rating` pada data `cc` . Bagaimana hubungan dan kekuatannya?

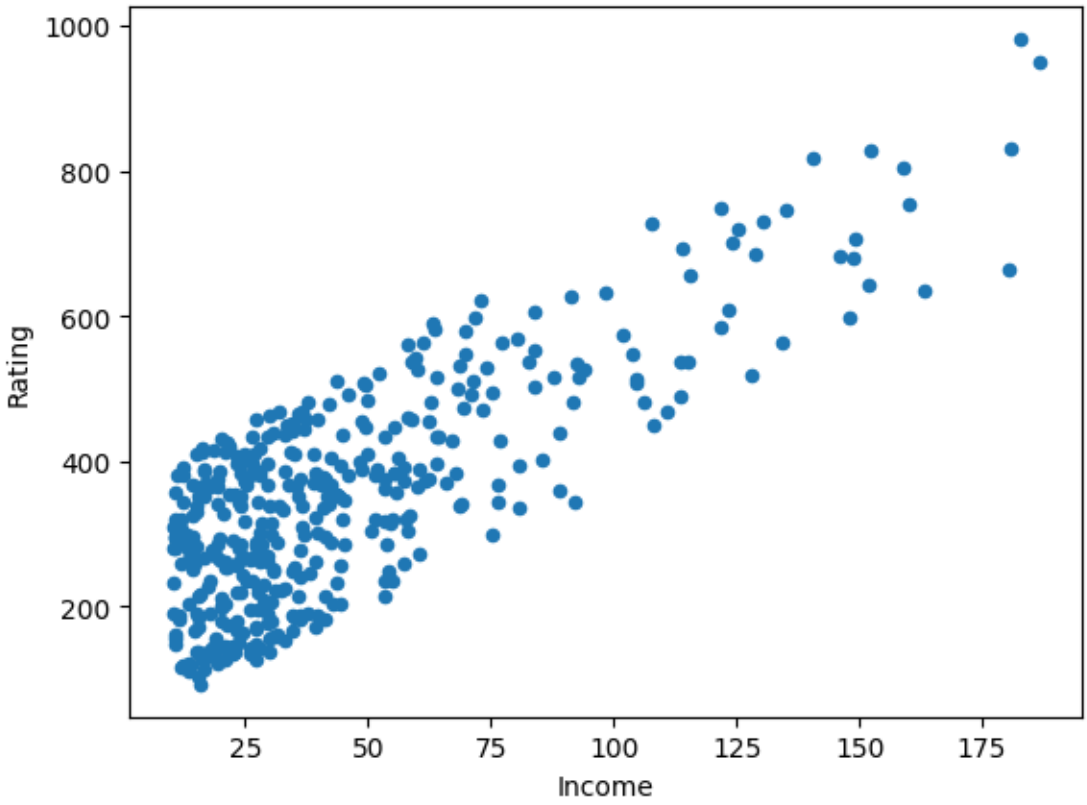
[ ]:
```
# code here


# end of code
```

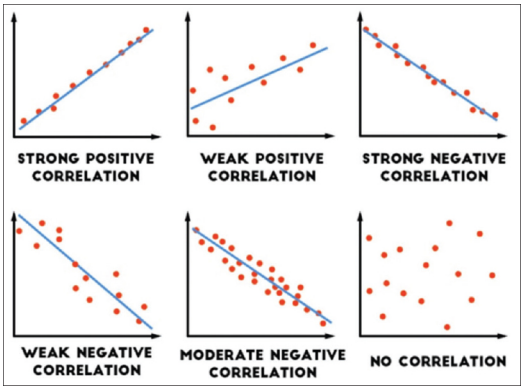
Jawaban: _____

Visualisasi korelasi dengan scatter plot:

[25]:
```
cc.plot.scatter(x='Income',
                y='Rating');
```



Ilustrasi correlation:



5.3.3 Knowledge Check

Dari pernyataan berikut, jawablah benar atau salah. Apabila salah, tuliskan pernyataan yang benar.

1. Ketika korelasi variabel A dan B bernilai -1 artinya tidak ada korelasi antara nilai A dan B.

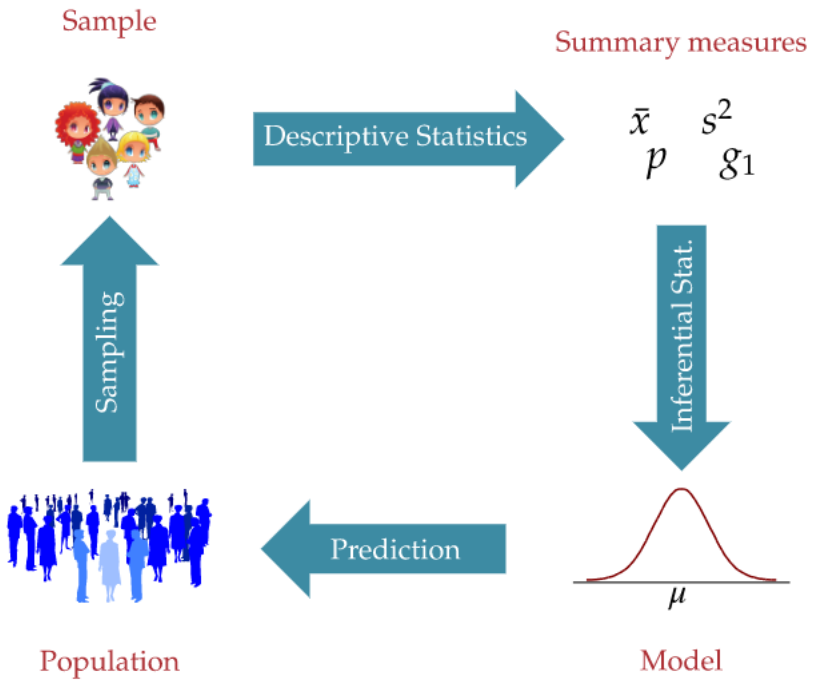
☐ Benar
☐ Salah

2. Scatter plot dapat digunakan untuk menggambarkan hubungan antara dua variabel numerik.

☐ Benar
☐ Salah

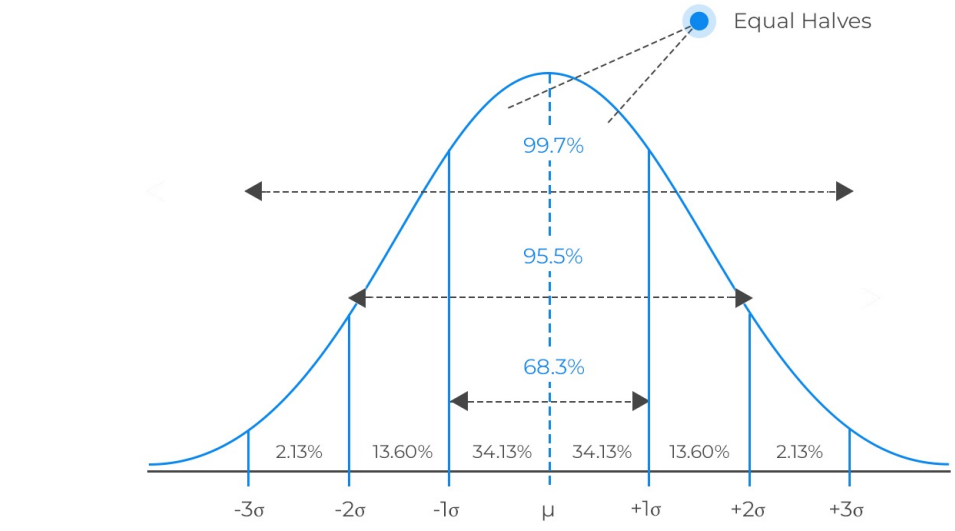
6 Inferential Statistics

Inferential Statistics membantu kita **menarik kesimpulan tentang keseluruhan data (populasi) dengan menggunakan sebagian informasinya saja (sampel)**



Setiap data memiliki distribusi. Distribusi data yang spresial dan berperan dalam inferential statistics adalah **distribusi normal**

6.1 Normal Distribution



Source: Quantitative Methods – Learning Sessions (Analyst Prep, 2019)

Karakteristik:

- Kurva membentuk lonceng simetris, artinya puncaknya adalah titik pusat (mean = median)
- Luas area dibawah kurva = 1 (menyatakan probabilitas)

- Persebaran data:
  - 68% data berada di rentang +- 1 standar deviasi dari mean
  - 95% data berada di rentang +- 2 standar deviasi dari mean
  - 99.7% data berada di rentang +- 3 standar deviasi dari mean
- **Standar normal baku** adalah distribusi normal dimana mean = 0 dan standar deviasi = 1.

Distribusi normal banyak digunakan pada inferensial statistik karena dicetuskannya **Central Limit Theorem**.

> Semakin bertambahnya jumlah sampel yang diambil secara acak, maka **distribusi rata-rata sampel** akan mengikuti distribusi normal

Karakteristik distribusi normal inilah yang dimanfaatkan untuk penghitungan inferensial statistik:

- **Menghitung Probabilitas:**
  - Probability Mass Function -> diskrit/kategorik
  - Probability Density Function -> kontinu/numerik
- **Membuat Confidence Interval**
- **Uji Hipotesis**

6.2 Probability Mass Function

- Menghitung peluang untuk data diskrit, contoh:
  - peluang hujan/tidak hujan
  - peluang produk yang terjual
  - peluang nasabah good credit/bad credit
- Formula: jumlah kejadian terjadi dibagi dengan jumlah kejadian total

Contoh:

Terdapat 100 nasabah dari sebuah Bank, 90 diantaranya merupakan nasabah dengan status good (good credit), sedangkan sisanya sebanyak 90 adalah status bad (bad credit). Berapakah peluang nasabah bad credit?

[ ]:
```
# code here


# end of code
```

6.3 Probability Density Function

- Menghitung probability data **kontinu**. Data kontinu merupakan data yang memiliki nilai dalam rentang tertentu, dan bisa memiliki angka desimal atau pecahan, contohnya:
  - tinggi badan
  - rating nasabah
  - profit/revenue
- Tahapan:
  1. Hitung Z-score (ubah nilai data asli ke standar normal baku = Z-score standardization)
  2. hitung peluang berdasarkan Z-score dengan menggunakan fungsi `pnorm()`
- Formula Z-score:

$$Z = \frac{x - \mu}{\sigma}$$

Keterangan:

- Z = Z-score
- x = titik data
- μ = mean
- σ = standar deviasi

  Z-score merupakan sebuah nilai yang merepresentasikan **berapa standard deviasi data tersebut menyimpang dari rata-ratanya**

Contoh

Tinggi badan pria dewasa di Indonesia berdistribusi normal dengan rata-rata 165 cm dan standar deviasi 10 cm. Berapa peluang pria dewasa di Indonesia memiliki tinggi badan > 180 cm?

Diketahui:

- mean = 165
- stdev = 10
- titik data = 180cm

[28]:
```
mean = 165
std = 10
titik_data = 180
```

[ ]:
```
# code here
from scipy.stats import norm

# hitung Z-score lalu ubah jadi peluang
Z_score = ....
```

[ ]:
```
# menghitung peluang
peluang = ....
```

Insight: Peluang pria dewasa di Indonesia memiliki tinggi badan > 180 cm _____

6.4 Confidence Interval

Confidence interval (selang kepercayaan) berguna untuk menduga nilai mean populasi dengan sebuah interval. Menebak dengan sebuah interval akan meminimalisir error dibandingkan hanya dengan menebak satu nilai.

- Formula:

$$CI = \bar{x} \pm Z_{\frac{\alpha}{2}} * SE$$

- Keterangan:
  - $\bar{x}$ = rata-rata sampel
  - $Z_{\frac{\alpha}{2}}$ = Z-score ketika alpha/2

  - α = tingkat error yang ditolerasi
  - tingkat kepercayaan = 1-α
  - SE = standard error

SE mengukur kebaikan sampel dalam mewakilkan populasi. Semakin kecil, maka sampel semakin representatif (baik).

$$SE = \frac{\sigma}{\sqrt{n}}$$

- Ket:
  - σ = standar deviasi populasi
  - n = jumlah sampel
- Tahapan:
  - hitung mean sampel
  - hitung standar deviasi sampel & SE
  - tentukan tingkat kepercayaan & α
  - tentukan Z alpha/2
  - hitung confidence interval

Contoh

Dari data `cc` yang berisikan sampel **400 nasabah** suatu Bank diketahui memiliki rata-rata `Balance` kredit sebesar **520**. Semisal diketahui Bank tersebut memiliki standard deviasi populasi untuk Balance sebesar **465**.

Berapakah confidence interval untuk rata-rata Balance seluruh nasabah? Gunakan tingkat kepercayaan 95%!

1. Diketahui:

- mean sampel = _____
- stdev populasi = _____
- jumlah sampel (n) = _____

2. Hitung nilai SE

[ ]:
```
# code here
import math
from scipy import stats

# std populasi dibagi akar n
SE = ...
```

2. Tentukan tingkat kepercayaan dan alpha

- Tingkat kepercayaan: 95%
- alpha (tingkat error): 100% - 95% = 5%, artinya kita mentoleransi error sebesar 5%, bahwa mungkin saja rata-rata Balance nasabah aslinya terletak di luar Confidence Interval

[ ]:
```
alpha =

print('alpha: ', alpha)
```

3. Hitung Z alpha/2

alpha dibagi 2 karena ingin membuat batas bawah dan batas atas (dalam dunia statistika dikenal sebagai two-tailed)

[ ]:
```
# code here
from scipy.stats import norm

# luas di bawah kurva - kedua bagian hijau
Z = ....
```

4. Hitung confidence interval

```
CI = mean -+ (Z * SE)
```

[ ]:
```
lower = mean - (Z*SE)
upper = mean + (Z*SE)
```

Kesimpulan: _____

6.5 Hypothesis Testing

Uji hipotesis bertujuan untuk menguji **dugaan**. Uji hipotesis sering disebut juga sebagai **uji signifikansi** yang digunakan untuk menguji apakah suatu treatment memberikan perubahan/pengaruh signifikan terhadap suatu kondisi.

Istilah-istilah:

- Hipotesis: dugaan sementara yang harus diuji
  - H_0 / null hypothesis:
    - kondisi awal
    - memiliki unsur kesamaan (=, >=, <=)
  - H_1 / alternative hypothesis:
    - kontradiktif dengan H_0
- α:
  - tingkat signifikansi yaitu tingkat error yang masih bisa ditoleransi
  - umumnya 0.05
- $1 - \alpha$: tingkat kepercayaan
- $p - value$:
  - hasil perhitungan statistik yang menunjukkan peluang data sampel terjadi dengan kondisi H0.

Pengambilan kesimpulan:

- Jika $p - value < \alpha$, maka tolak H_0 -> terima h1

- Jika $p - value > \alpha$, maka gagal tolak H_0 -> terima h0

Contoh Hipotesis

1. Hipotesis dua arah (!=)

- H_0 : Rata-rata saldo rekening tidak berbeda secara signifikan antara nasabah yang menggunakan layanan internet banking dan yang tidak menggunakan layanan tersebut. (=)
- H_1 : Rata-rata saldo rekening **berbeda secara signifikan** antara nasabah yang menggunakan layanan internet banking dan yang tidak menggunakan layanan tersebut. (!=)

2. Hipotesis satu arah (<)

- H_0 : Penambahan teller tidak memberikan perbedaan durasi pembayaran (>=)
- H_1 : Penambahan teller **menurunkan** durasi pembayaran (<)

3. Hipotesis satu arah (>)

- H_0: Penerapan diskon tidak memberikan perbedaan jumlah pembelian produk (<=)
- H_1: Penerapan diskon **meningkatkan** jumlah pembelian produk (>)

6.5.1 Z-Test

Uji hipotesis yang menggunakan Z-test bila:

- standar deviasi populasi diketahui
- jumlah sampel banyak (n > 30)

Contoh

BRI merupakan salah satu Bank terbaik di Indonesia. Bila diketahui rata-rata likes dari suatu post di platform mereka sebesar **14000** likes dengan standar deviasi **5000** likes.

Demi meningkatkan likes dari tiap post, BRI memutuskan untuk menggunakan influencer sebagai brand ambassador pemasaran produk. Setelah menggunakan influencer, diambil **50** postingan acak yang ternyata memiliki rata-rata likes **17500**.

Sebagai tim marketing, lakukan analisis apakah menggunakan jasa influencer secara signifikan meningkatkan customer engagement (dari sisi rata-rata jumlah likes) atau tidak? Gunakan tingkat kepercayaan **95%**.

Jawaban:

I. Tentukan hipotesis

- H_0:
- H_1:

II. Hitung nilai statistik

Diketahui deskriptif statistiknya:

- mean populasi =
- stdev populasi =
- n =
- mean sampel =

Ditentukan oleh user:

- tingkat kepercayaan =
- alpha =

[ ]:
```
# nilai statistic descriptive
mean_populasi =
std_populasi =
n =
mean_sample =

print('mean_populasi: ', mean_populasi)
print('std_populasi: ', std_populasi)
print('n: ', n)
print('mean_sample: ', mean_sample)
```

$$Z = \frac{\bar{X} - \mu}{SE}$$

Z = (rata2 sampel - rata2 populasi) / standar error

$$SE = \frac{\sigma}{\sqrt{n}}$$

Standar error = standar deviasi populasi / akar dari banyak sampel

[ ]:
```
# menghitung nilai SE
SE =

# menghitung nilai z
Z =
```

[ ]:
```
p_value =
```

Gunakan fungsi `ztest()` untuk menghitung z-statistics dan p-value.

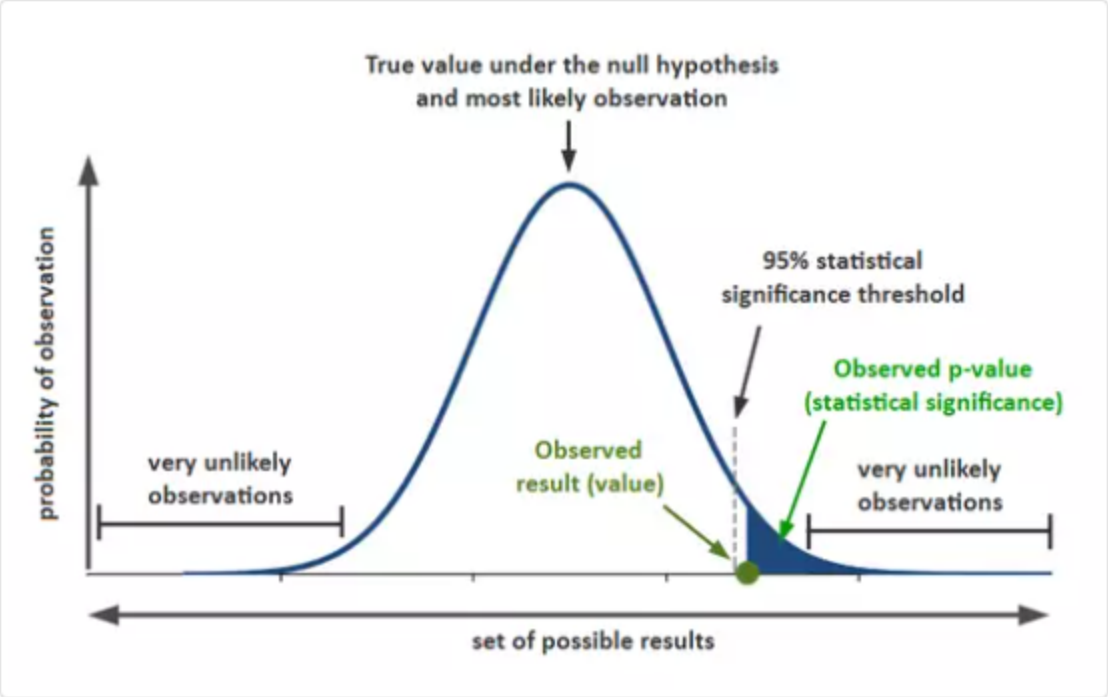
```
zstats, pval = ztest(x1=...,
                     value = ...,
                     alternative = ...)
```

parameter:

- `x1` : number of observations
- `value` : rata-rata dari x1 di H_0
- `alternative` :
  - jika H_1 tidak sama (!=) dengan nilai tertentu, isi dengan `two-sided`
  - jika H_1 lebih besar (>) dari suatu nilai, gunakan `larger`
  - jika H_1 lebih kecil (<) dari suatu nilai, gunakan `smaller`

```
[ ]: from statsmodels.stats.weightstats import ztest

z_test =
```



c. Bandingkan P-value dengan alpha

Pengambilan kesimpulan:

- Jika $p - value < \alpha$, maka tolak H_0
- Jika $p - value > \alpha$, maka gagal tolak H_0

p-value = _____ (>/<) alpha = 0.05

IV. Kesimpulan

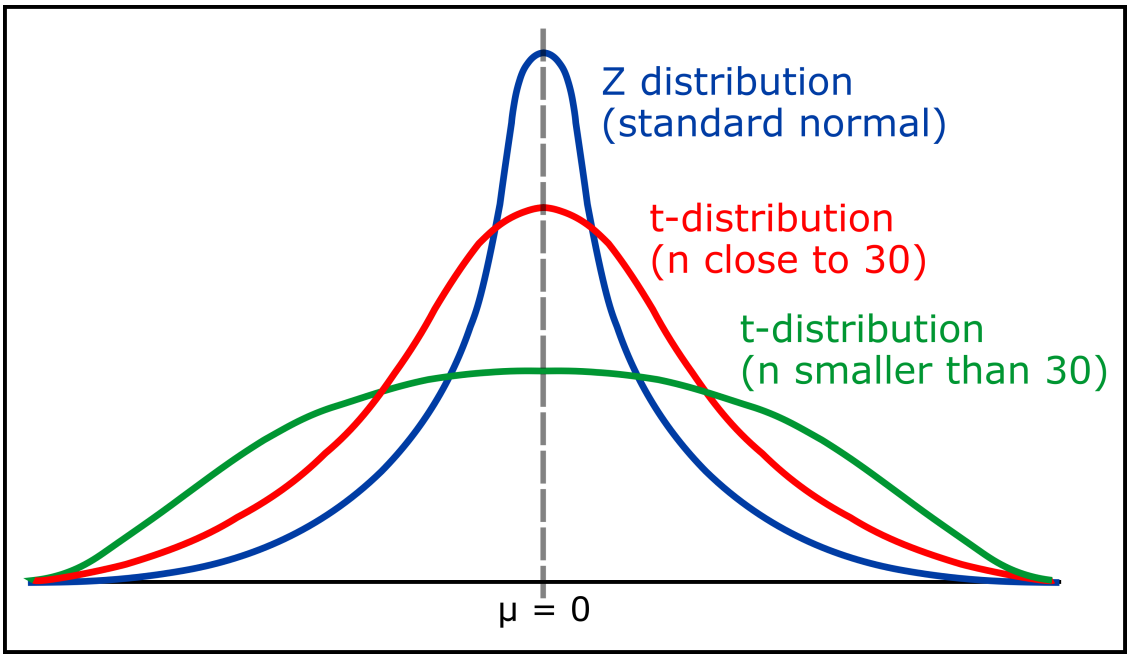
...

6.5.2 T-test

Uji hipotesis menggunakan T-test jika:

- standar deviasi populasi tidak diketahui atau
- jumlah sampel sedikit (n <= 30)

Bentuk t-distribution mirip dengan normal distribution, hanya saja lebih landai ketika jumlah sampel sedikit:



Contoh Kasus

Mari kita asumsikan Bank BRI memiliki dua kelompok nasabah bank, yaitu kelompok yang memiliki behavior scoring tinggi dan kelompok yang memiliki behavior scoring rendah.

Diketahui data saldo rekening antara kedua kelompok sebagai berikut:

[30]:
```
behavior_score_high = pd.Series([30.4, 52.7, 70.6, 55.7, 56.3, 34.2, 59.6, 42.
↪3, 21.1, 50.5, 12.2, 58.6, 12.0, 56.1, 49.4, 60.9, 60.0, 35.3, 15.0, 50.3])

behavior_score_low = pd.Series([6.5, 13.3, 6.8, 9.2, 10.0, 1.5, 21.7, 16.2, 5.
↪9, 25.0, 18.4, 12.6, 22.2, 22.0, 21.6, 20.5, 19.4, 14.5, 12.6, 12.0])
```

Tujuan kita adalah menguji **apakah terdapat perbedaan signifikan dalam rata-rata saldo rekening antara kedua kelompok tersebut?**

Jawab:

I. Tentukan hipotesis

- H_0: Rata-rata saldo rekening antara kelompok dengan behavior scoring tinggi dan kelompok dengan behavior scoring rendah tidak berbeda secara signifikan.
- H_1: Terdapat perbedaan signifikan dalam rata-rata saldo rekening antara kedua kelompok.

II. Hitung P-value dengan `ttest_ind()`

Gunakan fungsi `ttest_ind()` untuk menghitung t-statistics dan p-value dua independent sample

```
t_stats, pval = ttest_ind(a= ...,
                          b= ...,
                          alternative = ...)
```

parameter:

- `a` : data atau observasi sampel berbentuk Series atau array
- `b` : data atau observasi sampel berbentuk Series atau array
- `alternative` : tergantung hypothesis alternative (H_1)
  - jika H_1 tidak sama (!=) dengan nilai tertentu, isi dengan `two-sided`
  - jika H_1 lebih besar (>) dengan nilai tertentu, isi dengan `less`
  - Jika H_1 lebih besar (<) dengan nilai tertentu, isi dengan `greater`

[ ]:
```
# code here
from scipy import stats

t_test =
```

III. Bandingkan P-value dengan alpha

Dalam membuat keputusan uji statistik, kita dapat membandingan p-value dengan alpha:

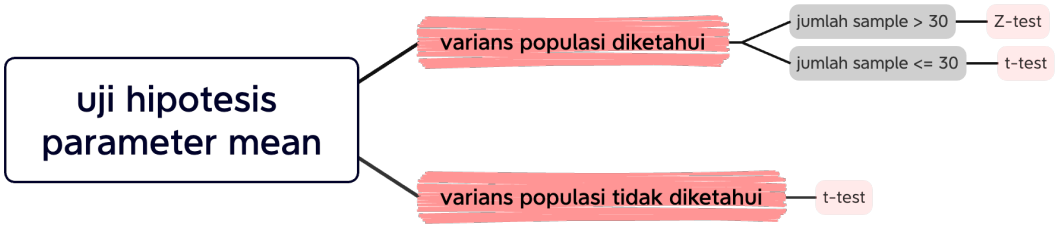
- selang kepercayaan = 95%
- alpha = 5% -> 0.05

p-value = _____ (>/<) alpha = 0.05, maka _____

IV. Kesimpulan

Dengan menggunakan tingkat kepercayaan 95% dapat disimpulkan _____

Summary penggunaan hipotesis testing:



7 Further Readings

- Descriptive Statistics: https://courses.lumenlearning.com/suny-natural-resources-biometrics/chapter/chapter-1-descriptive-statistics-and-the-normal-distribution/
- Dealing with small data set: https://measuringu.com/small-n/
- t-Distribution and some case examples: https://stattrek.com/probability-distributions/t-distribution.aspx