Rakamin Academy

Data Scientist
at Telecommunication Company

Halo!

Perkenalkan saya **Abdul**lah Ghifari.

Sebelumnya

bukalapak

tiket.com

Institut Pertanian Bogor

Abdullah Ghifari
https://www.linkedin.com/in/abdullah-ghifari/

Rakamin Academy

Statistics

Introduction to Statistics

Outline Pembelajaran

Rakamin Academy

Introduction to Statistics

- Introduction to Statistics
- Data Types
- Descriptive Statistics
- Measure of Central Tendency
- Measure of Spread
- Hands-On

Objektif

Memahami apa itu statistika, tipe tipe data dalam statistik (kategori dan numerik), Descriptive Analysis beserta prinsip dan implementasinya

Expected Output

Memahami apa itu statistika
Memahami tipe tipe data dalam statistik (kategori dan numerik)
Memahami apa itu Descriptive Statistics beserta prinsipnya
Mampu melakukan Descriptive Statistics dan kapan diimplementasikannya

Hands-On Required :

Hands - On : Statistics I.ipynb

Dataset :

1. **HR_comma_sep.csv**

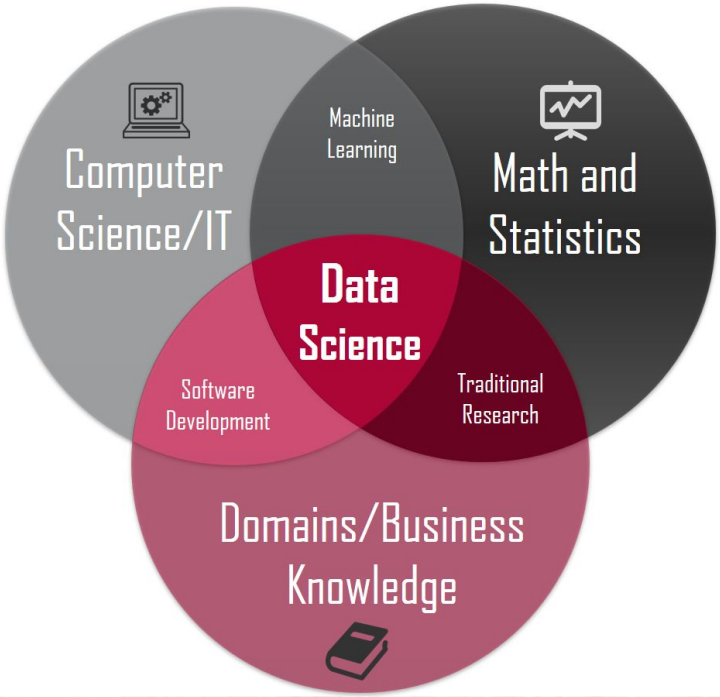
Klik disini untuk mengakses folder Hands-On dan Dataset

Outline Pembelajaran

Rakamin Academy

Introduction to Statistics

- Introduction to Statistics
- Data Types
- Descriptive Statistics
- Measure of Central Tendency
- Measure of Spread
- Hands-On



A **Data scientist** is one who knows more statistics than a programmer and more programming than a statistician.

Berapa persen pengguna sepeda motor di Jakarta?

Macam macam pertanyaan bisnis yang bisa dijawab oleh statistika

1. Bagaimana tingkat kepuasan customer jika menggunakan model A atau model B?
2. Apa faktor yang menyebabkan user menjadi churn?
3. Apakah product mengalami peningkatan yang signifikan ataukah sebaliknya?

Definisi Statistika

Disiplin Ilmu yang mempelajari teknik **pengumpulan, pengolahan, analisis, interpretasi dan presentasi** data

Definisi Statistik

Nanti akan kita bahas perbedaannya ya! ^^


Data
Statistika
Kesimpulan

Bagaimana caranya menarik kesimpulan?

Statistik Deskriptif
Mengambil kesimpulan dari sampel, menggunakan mean, median dan modus dan beberapa visualisasi

Statistika Inferensi
Mengambil kesimpulan untuk populasi menggunakan uji hipotesis

Populasi dan Sampel

Populasi

Totalitas dari semua objek yang hendak diteliti
(Contoh : Seluruh Masyarakat Jakarta)

Sampel

bagian kecil dari populasi yang diambil sebagai objek pengamatan karena dianggap bisa mewakili populasi
(Contoh : beberapa orang yang terpilih pada masing-masing provinsi)

Populasi dan Sampel

Raka pergi membeli buah lengkeng di toko buah. Kumpulan buah lengkeng ditumpuk pada sebuah rak. Untuk memastikan rasa dari buah lengkeng tersebut,

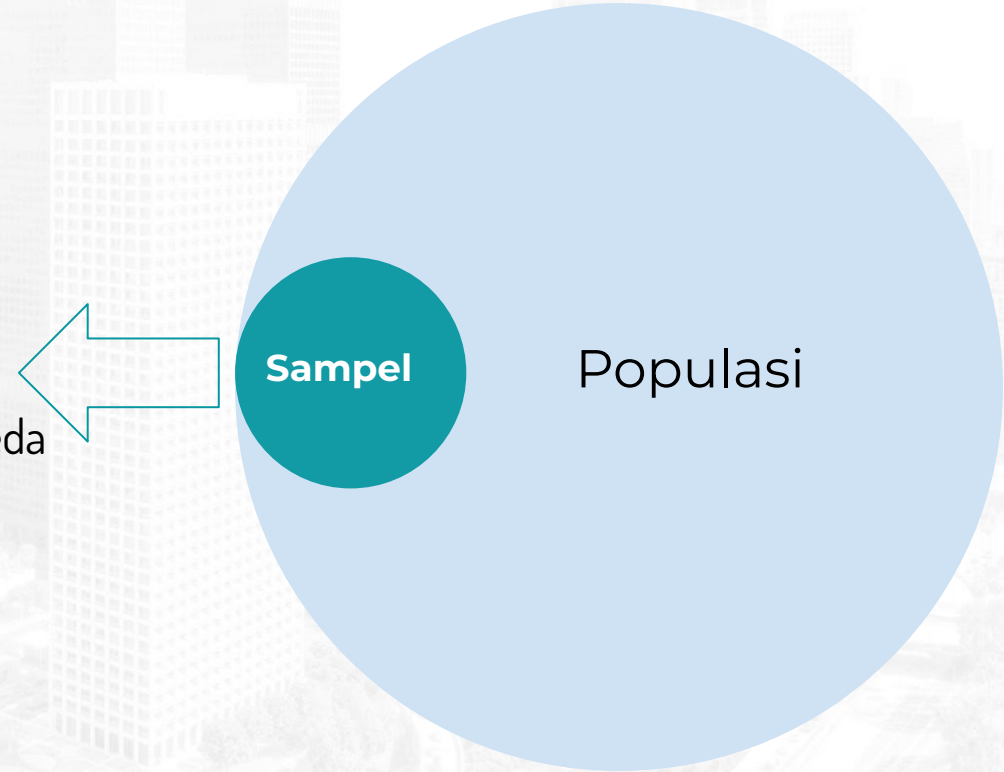
Raka kemudian mengambil beberapa buah lengkeng dan dicoba langsung. Karena buah lengkeng yang dicobanya manis, maka **Raka** pun membeli buah di toko tersebut.

Populasi dan Sampel

Populasi : Populasi pada ilustrasi di atas adalah seluruh buah lengkeng yang ada pada rak di toko buah tadi.

Sampel : Buah yang Raka ambil untuk dicicipi dengan tujuan untuk menentukan rasa manisnya.

Sampel diambil hanya sekian % dari populasi, populasi dan sampel menghasilkan ***karakteristik*** yang berbeda



Karakteristik

Populasi

Disebut dengan **parameter**.

Contoh : rata-rata dari seluruh siswa kelas 8 adalah 70

Sampel

Disebut dengan **statistik**.

Contoh : rata-rata dari 5 siswa kelas 8 adalah 65

Outline Pembelajaran

Rakamin Academy

Introduction to Statistics

- Introduction to Statistics
- Data Types
- Descriptive Statistics
- Measure of Central Tendency
- Measure of Spread
- Hands-On

Apa itu Data?

Data

Satuan unit informasi

Disusun dalam bentuk data matrix :

- Baris : menggambarkan observation
- Kolom : menggambarkan variabel (peubah)

Data Matrix

Observation

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	division	salary
0	0.38	0.53	2	157	3	0	1	0	sales	low
1	0.80	0.86	5	262	6	0	1	0	sales	medium
2	0.11	0.88	7	272	4	0	1	0	sales	medium
3	0.72	0.87	5	223	5	0	1	0	sales	low
4	0.37	0.52	2	159	3	0	1	0	sales	low

Variable

Kategorik

1. **Nominal**
   Skala nominal biasa digunakan untuk tipe peubah yang bersifat kategori dan tidak bisa diurutkan (Ex : Pria-Wanita, Warna, Pekerjaan)

2. **Ordinal**
   Skala Ordinal biasa digunakan untuk tipe peubah yang bersifat kategori dan bisa diurutkan tapi tidak dihitung besarnya perbedaan. (Ex : Suka-Tidak Suka, Jenjang Sekolah,)

Numerik

1. **Diskret**
   Skala diskret merepresentasikan barang yang dapat dihitung dengan dicacah. Seperti 0, 1, 2, 3 (Dapat dihitung). (Ex : Jumlah murid, Jumlah kendaraan)

2. **Kontinu**
   Skala kontinu merepresentasikan pengukuran, angkanya tidak dapat dihitung secara dicacah biasa digambarkan dengan cara interval. (Ex : Tinggi, Temperatur, Kecepatan)

Tipe Data Dari Tabel berikut

	satisfaction_level	last_evaluation	number_project	average_montly_hours	time_spend_company	Work_accident	left	promotion_last_5years	division	salary
0	0.38	0.53	2	157	3	0	1	0	sales	low
1	0.80	0.86	5	262	6	0	1	0	sales	medium
2	0.11	0.88	7	272	4	0	1	0	sales	medium
3	0.72	0.87	5	223	5	0	1	0	sales	low
4	0.37	0.52	2	159	3	0	1	0	sales	low

Outline Pembelajaran

Rakamin Academy

Introduction to Statistics

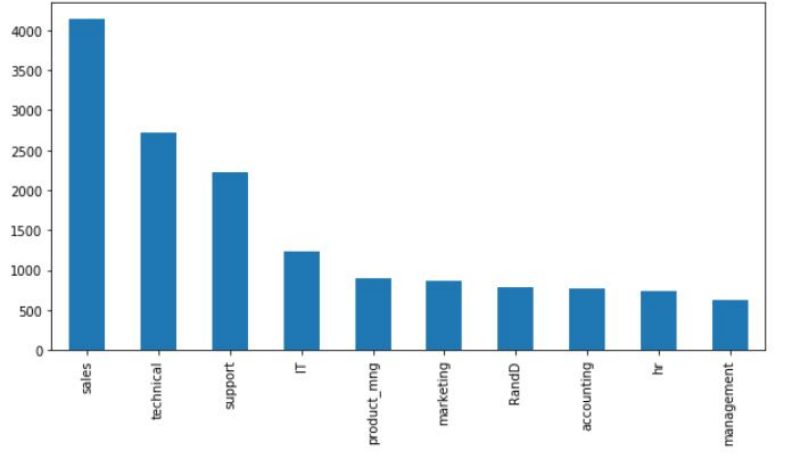
- [x] Introduction to Statistics
- [x] Data Types
- [ ] Descriptive Statistics
- [ ] Measure of Central Tendency
- [ ] Measure of Spread
- [ ] Hands-On

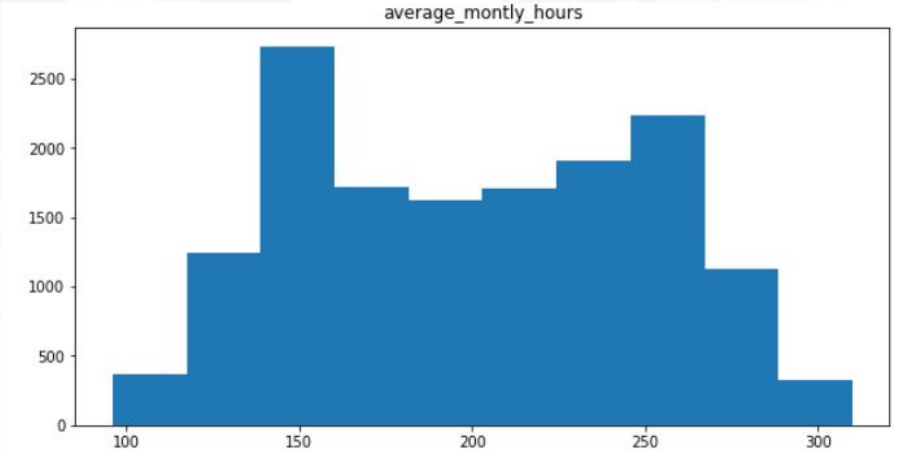
Statistika Deskriptif

Langkah pertama dalam menganalisis sekumpulan data untuk memiliki gagasan yang bagus tentang seperti apa data itu. Ini adalah fungsi ***statistik deskriptif***.

Salah satu *tools* dasar untuk menggambarkan distribusi nilai untuk beberapa variabel dalam sampel subjek adalah **histogram/barchart**.

Sebaran (Distribution)





Data Kategorik/Diskret

Grafik batang yang menunjukkan berapa banyak pengamatan yang termasuk dalam setiap kategori

Data Kontinu

Bagilah rentang nilai menjadi beberapa interval yang pas, hitung berapa banyak pengamatan yang termasuk dalam setiap interval, dan kemudian tampilkan jumlah itu dalam diagram batang.

Statistika Deskriptif

Measure of **Central Tendency**
(Ukuran Pemusatan)

Measure of **Spread**
(Ukuran Penyebaran)

Statistika Deskriptif

Measure of **Central Tendency** :

1. Mean (Rata-rata)
2. Median
3. Mode (Modus)

Outline Pembelajaran

Rakamin Academy

Introduction to Statistics

- [x] Introduction to Statistics
- [x] Data Types
- [x] Descriptive Statistics
- [ ] Measure of Central Tendency
- [ ] Measure of Spread
- [ ] Hands-On

Statistika Deskriptif

Mode

Modus adalah data yang memiliki jumlah frekuensi paling tinggi

Statistika Deskriptif

Contoh

Kita memiliki list of data sebagai berikut

3,3,4,4,4,4,4,5,5,5,6,7,7,7,7,7,7,9,9,10

Carilah **mode** dari data berikut

Statistika Deskriptif

Mode

Modus adalah data yang memiliki jumlah frekuensi paling tinggi

Contoh :
Mode = 7

Data	Frequency
3	2
4	5
5	3
6	1
7	6
8	0
9	2
10	1
Total	20

Statistika Deskriptif

Mean

Rata-rata adalah nilai tengah. Penjumlahan dari setiap nilai dibagi dengan banyaknya data.

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

Statistika Deskriptif

Contoh

Kita memiliki list of data sebagai berikut

3,3,4,4,4,4,4,5,5,5,6,7,7,7,7,7,7,9,9,10

Carilah **mean** dari data berikut

Statistika Deskriptif

mean =

Statistika Deskriptif

mean = (3+3+4+4+4+4+4+5+5+5
+6+7+7+7+7+7+7+9+9+10)/20

Statistika Deskriptif

mean = (3+3+4+4+4+4+4+5+5+5
+6+7+7+7+7+7+7+9+9+10)/20
= 117 / 20

Statistika Deskriptif

mean = (3+3+4+4+4+4+4+5+5+5
+6+7+7+7+7+7+7+9+9+10)/20
= 117 / 20
= 5.85

Statistika Deskriptif

Sifat Rata-rata (Mean)

Rata-rata sangat terpengaruh oleh nilai yang sangat besar atau kecil.

Nilai ini biasa disebut dengan **outliers**.

Tidak disarankan untuk menggunakan rata-rata apabila terdapat outliers pada data kita. **(not robust)**

Contoh Kasus

Kita bekerja pada perusahaan e-commerce untuk tim seller management. Stakeholder meminta tolong untuk memeriksa performa **transaksi perhari** dari masing-masing seller.

Diperoleh data sebagai berikut

	Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Average
Data 1	10	20	30	40	50	30

Contoh Kasus

Kita bekerja pada perusahaan e-commerce untuk tim seller management. Stakeholder meminta tolong untuk memeriksa performa **transaksi perhari** dari masing-masing seller.

Diperoleh data sebagai berikut

	Seller 1	**Seller 2**	**Seller 3**	**Seller 4**	**Seller 5**	**Average**
Data 1	10	20	30	40	50	30
Data 2	10	20	30	40	***1000***	**220**

Contoh Kasus

	Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Average
Data 1	10	20	30	40	50	30
Data 2	10	20	30	40	***1000***	**220**

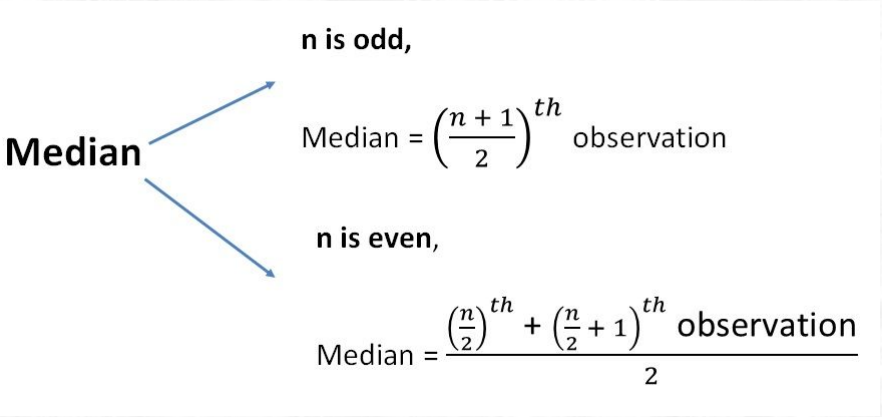
Jika stakeholder bertanya, kira-kira mana aja seller yang penjualannya per hari nya mendekati nilai rata-rata.

Data 1: Seller 2, Seller 3, dan Seller 4
Data 2: Tidak ada

Statistika Deskriptif

Median

Titik tengah data yang terletak 50% dari keseluruhan data.



Statistika Deskriptif

Contoh

Kita memiliki list of data sebagai berikut

3,3,4,4,4,4,4,5,5,5,6,7,7,7,7,7,7,9,9,10

Carilah **median** dari data berikut

Statistika Deskriptif

1. Tentukan formula ganjil/genap

Data	Frekuensi	Frekuensi Kumulatif
3	2	2
4	5	7
5	3	10
6	1	11
7	6	17
8	0	17
9	2	19
10	1	20
Total	20	

Statistika Deskriptif

1. Tentukan formula ganjil/genap

   Genap

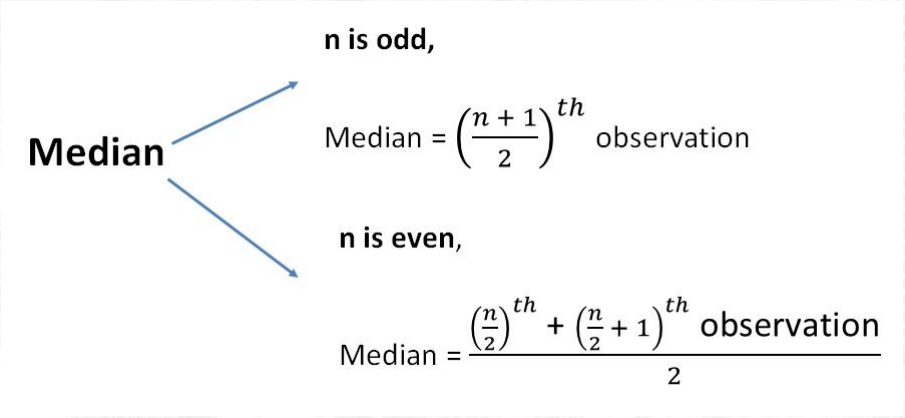
Data	Frekuensi	Frekuensi Kumulatif
3	2	2
4	5	7
5	3	10
6	1	11
7	6	17
8	0	17
9	2	19
10	1	20
Total	20	

Statistika Deskriptif

1. Tentukan formula ganjil/genap

   Genap

2. Tentukan letak median berada



Data	Frekuensi	Frekuensi Kumulatif
3	2	2
4	5	7
5	3	10
6	1	11
7	6	17
8	0	17
9	2	19
10	1	20
Total	20	

Statistika Deskriptif

1. Tentukan formula ganjil/genap

   Genap

2. Tentukan letak median berada

   Letak ke - 20/2 = letak ke -10 = 5

   Letak ke - 20/2 + 1 = letak ke - 11 = 6

Data	Frekuensi	Frekuensi Kumulatif
3	2	2
4	5	7
5	3	10
6	1	11
7	6	17
8	0	17
9	2	19
10	1	20
Total	20	

Statistika Deskriptif

1. Tentukan formula ganjil/genap
   Genap
2. Tentukan letak median berada
   Letak ke - 20/2 = letak ke -10 = 5
   Letak ke - 20/2 + 1 = letak ke - 11 = 6
3. Cari nilai median dengan menggunakan rumus tersebut
   Median = (5+6)/2 = 5.5

Data	Frekuensi	Frekuensi Kumulatif
3	2	2
4	5	7
5	3	10
6	1	11
7	6	17
8	0	17
9	2	19
10	1	20
Total	20	

Statistika Deskriptif

Sifat Median

Median relatif robust dari outlier (tidak terpengaruh oleh nilai yang sangat tinggi atau rendah)

Biasanya digunakan untuk distribusi skew (menceng)

Contoh Kasus

Kita bekerja pada perusahaan e-commerce untuk tim seller management. Stakeholder meminta tolong untuk memeriksa performa **transaksi perhari** dari masing-masing seller.

Diperoleh data sebagai berikut

	Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Median
Data 1	10	20	30	40	50	30

Contoh Kasus

Kita bekerja pada perusahaan e-commerce untuk tim seller management. Stakeholder meminta tolong untuk memeriksa performa **transaksi perhari** dari masing-masing seller.

Diperoleh data sebagai berikut

	Seller 1	**Seller 2**	**Seller 3**	**Seller 4**	**Seller 5**	**Median**
Data 1	10	20	30	40	50	30
Data 2	10	20	30	40	***1000***	**30**

Contoh Kasus

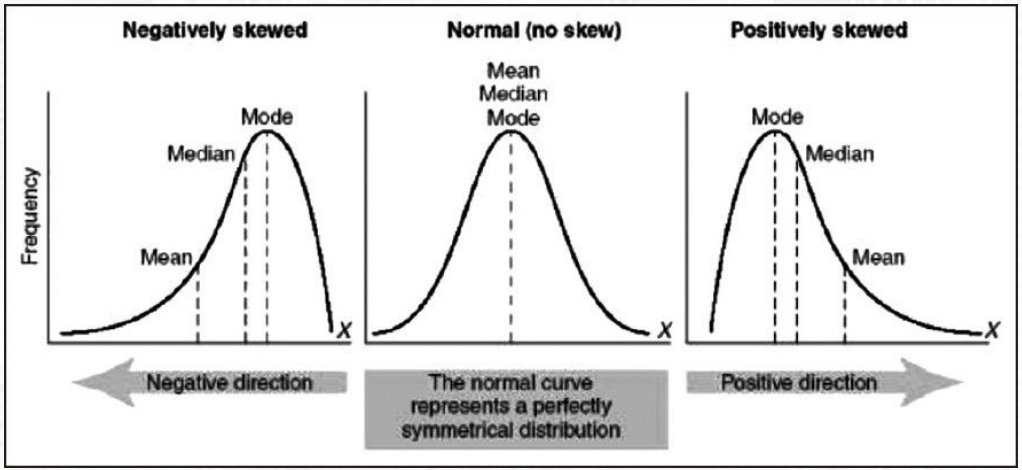
	Seller 1	Seller 2	Seller 3	Seller 4	Seller 5	Median
Data 1	10	20	30	40	50	30
Data 2	10	20	30	40	***1000***	**30**

Jika stakeholder bertanya, kira-kira mana aja seller yang penjualannya per hari nya mendekati nilai rata-rata.

Data 1: Seller 2, Seller 3, dan Seller 4
Data 2: Seller 2, Seller 3, dan Seller 4

Statistika Deskriptif

Tipe-tipe Distribusi



Negatively Skewed

1. Penilaian performa karyawan
2. Discount value claim distribution

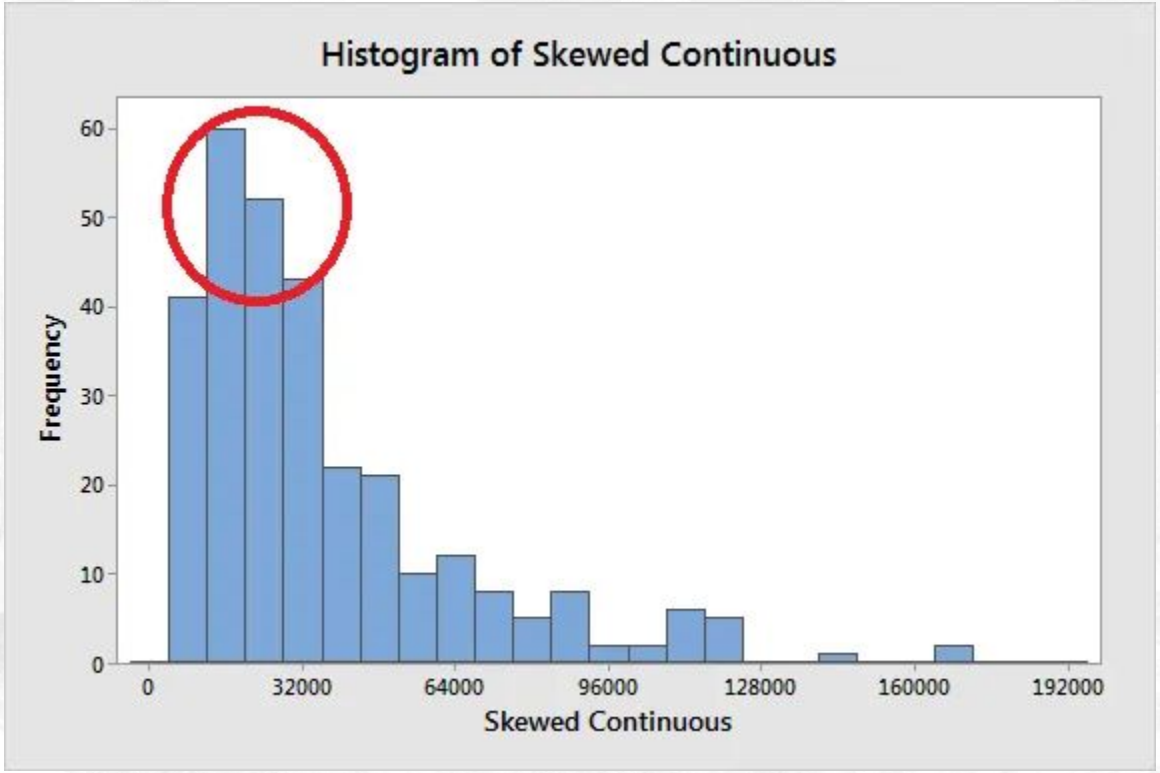
Normal

1. Durasi pengiriman makanan
2. Distribusi tinggi badan

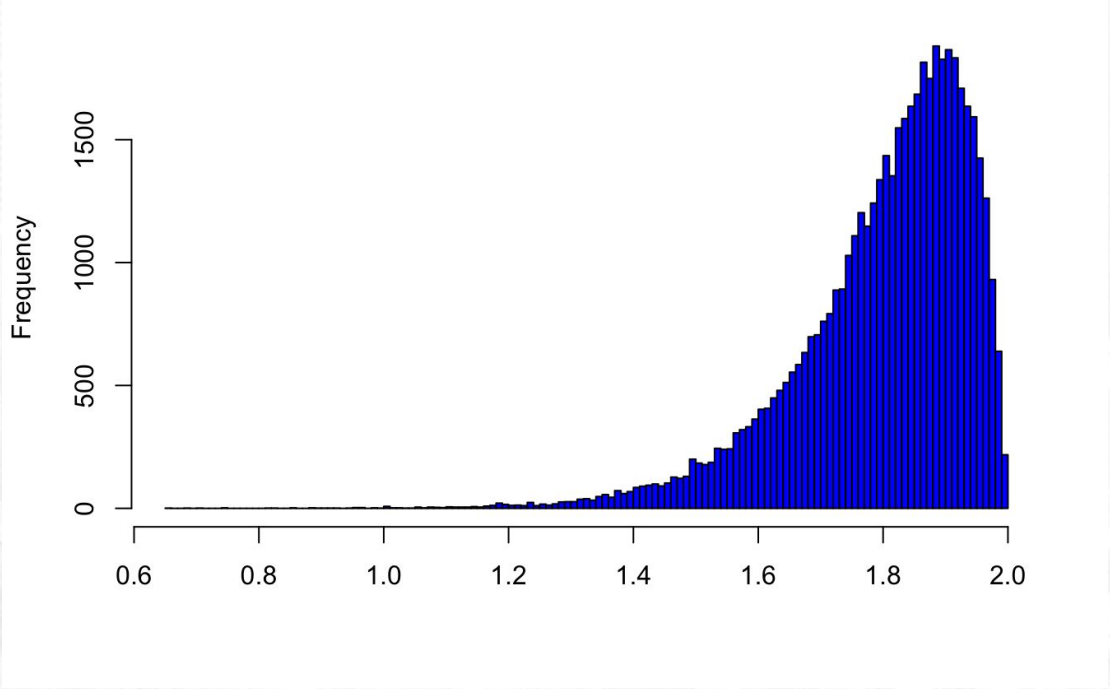
Positively Skewed

1. Income distribution
2. Complaint resolve time

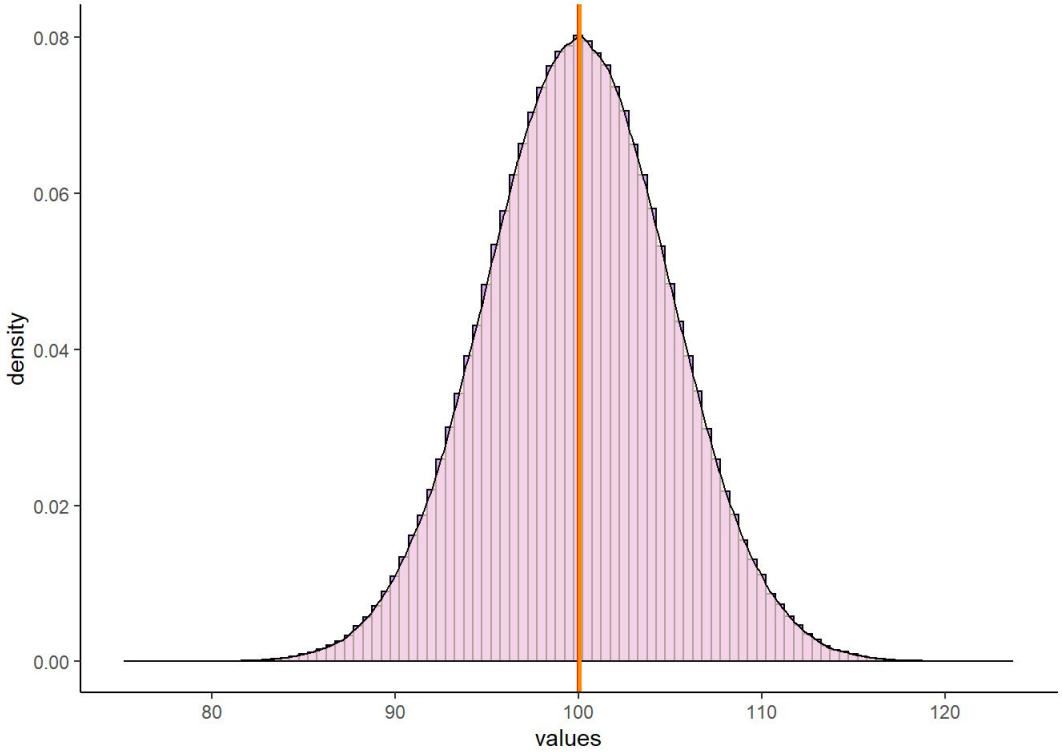
Termasuk distribusi apakah grafik ini?



Termasuk distribusi apakah grafik ini?



Termasuk distribusi apakah grafik ini?



Statistika Deskriptif

Ukuran Pemusatan Lainnya

Quartile : Dibagi menjadi 4 bagian. Biasa disimbolkan dengan Q

Quartile 1 (Q1) = Terletak pada 25% data
Quartile 2 (Q2) = Median (Titik tengah-Terletak pada 50% data)
Quartile 3 (Q3) = Terletak pada 75% data
Quartile 4 (Q4) = Terletak pada 100% data

Statistika Deskriptif

1. **Carilah Q1!**

Data	Frekuensi	Frekuensi Kumulatif
3	2	2
4	5	7
5	3	10
6	1	11
7	6	17
8	0	17
9	2	19
10	1	20
Total	20	

Statistika Deskriptif

1. Carilah Q1!

Letak ke - ¼ data = letak ke - 5 = 4

Data	Frekuensi	Frekuensi Kumulatif
3	2	2
4	5	7
5	3	10
6	1	11
7	6	17
8	0	17
9	2	19
10	1	20
Total	20	

Statistika Deskriptif

1. Carilah Q1!

Letak ke - ¼ data = letak ke - 5 = 4

2. Carilah Q2!

Q2 = Median = 5.5

Data	Frekuensi	Frekuensi Kumulatif
3	2	2
4	5	7
5	3	10
6	1	11
7	6	17
8	0	17
9	2	19
10	1	20
Total	20	

Statistika Deskriptif

1. Carilah Q1!

Letak ke - ¼ data = letak ke - 5 = 4

2. Carilah Q2!

Q2 = Median = 5.5

3. Carilah Q3!

Letak ke - ¾ data = letak ke - 15 = 7

Data	Frekuensi	Frekuensi Kumulatif
3	2	2
4	5	7
5	3	10
6	1	11
7	6	17
8	0	17
9	2	19
10	1	20
Total	20	

Statistika Deskriptif

1. Carilah Q1!

Letak ke - ¼ data = letak ke - 5 = 4

2. Carilah Q2!

Q2 = Median = 5.5

3. Carilah Q3!

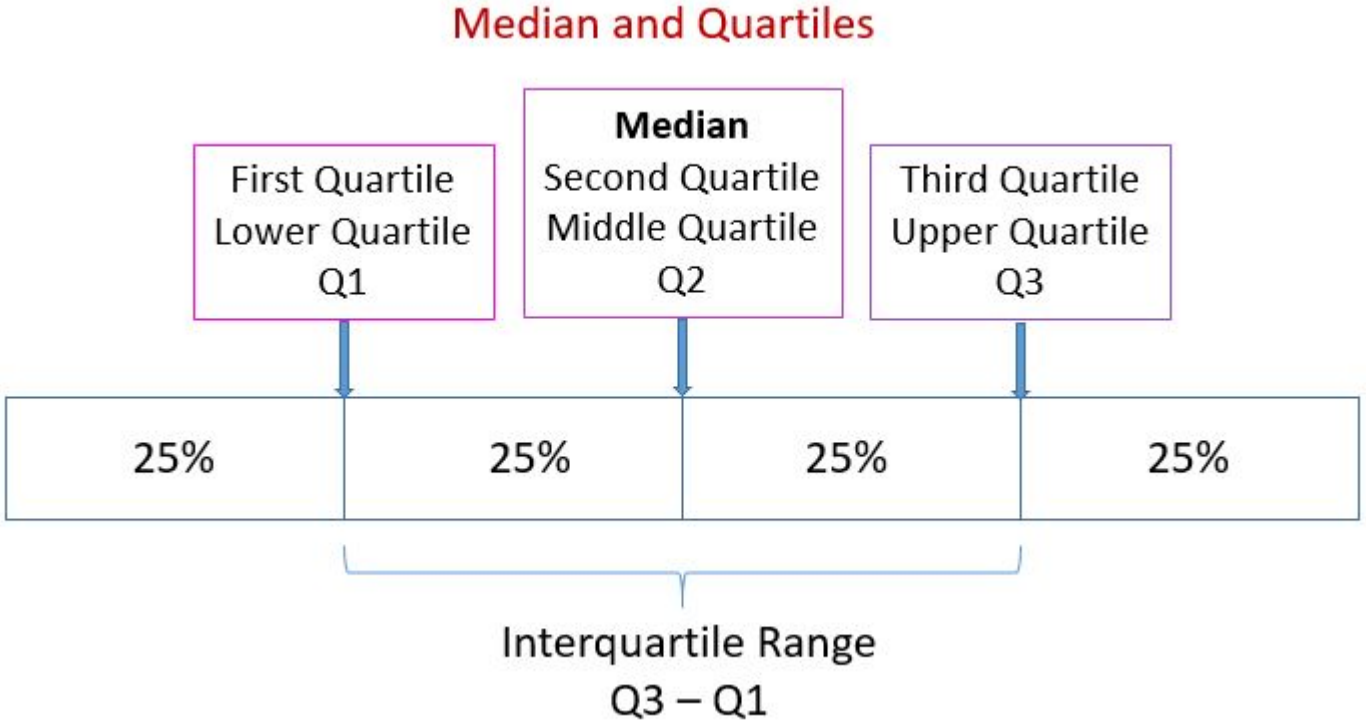
Letak ke - ¾ data = letak ke - 15 = 7

4. Carilah Q4!

Q4 = Max data = 10

Data	Frekuensi	Frekuensi Kumulatif
3	2	2
4	5	7
5	3	10
6	1	11
7	6	17
8	0	17
9	2	19
10	1	20
Total	20	

Statistika Deskriptif



Contoh Penerapan Quartil

Tim customer service mendapat pesan dari direksi agar mampu **menyelesaikan complaint maksimal 1 hari**.

Kira-kira bagaimana cara mengukur performa tim yang baik?

Apakah dengan menggunakan rata-rata?

Ataukah kita perlu menggunakan nilai maximum sesuai dengan perintah direksi?

Contoh Penerapan Quartil

Statistika Deskriptif

Ukuran Pemusatan Lainnya

Percentile : Persentase lokasi dari observasi yang telah diurutkan

Misalkan, kita mendapat angka percentile 20 sebesar **60**.

Maka 20% observasi berada di bawah **60**
Similar dengan
80% sisanya berada di atas **60**.

Percentile 25th = Q1
Percentile 50th = Q2
Percentile 75th = Q3

Penerapan Percentil

Percentile digunakan ketika:

1. Ingin mengukur dengan toleransi error yang custom
2. Ingin melihat distribusi secara custom

Contoh:

Sebuah perusahaan pengiriman pesan online (Messenger Online) memiliki target supaya pesan dikirim dan sampai ke penerima dalam waktu maksimal 1 detik. Tim berusaha seoptimal mungkin membuat sistem tersebut. Namun, tetap saja akan ada outlier yang melebihi 1 detik. Maka percentile dapat digunakan. Misal P95 atau P99.

Outline Pembelajaran

Rakamin Academy

Introduction to Statistics

- [x] Introduction to Statistics
- [x] Data Types
- [x] Descriptive Statistics
- [x] Measure of Central Tendency
- [ ] Measure of Spread
- [ ] Hands-On

Statistika Deskriptif

Measure of **Spread** (Ukuran Penyebaran) :

1. Range (Jangkauan)
2. Variance (Ragam)
3. Interquartile (Interkuartil)

Statistika Deskriptif

Range (Jangkauan)

Jarak antara nilai maksimum dan nilai minimum

Range = maximum value - minimum value

Data	Frequency
3	2
4	5
5	3
6	1
7	6
8	0
9	2
10	1
Total	20

Statistika Deskriptif

Range (Jangkauan)

Jarak antara nilai maksimum dan nilai minimum

Range = maximum value - minimum value

Range = 10-3 = 7

Data	Frequency
3	2
4	5
5	3
6	1
7	6
8	0
9	2
10	1
Total	20

Statistika Deskriptif

Variance (Ragam)

Rata-rata kuadrat selisih dari mean

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Standard Deviation (Simpangan Baku)

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

Data	Frequency
3	2
4	5
5	3
6	1
7	6
8	0
9	2
10	1
Total	20

Ilustrasi Simpangan Baku dan Ragam

Data	Rata-rata	Simpangan Baku (Standard Deviation)	Ragam (Variance)
6, 6, 8, 8	7	1	1
0, 6, 8, 14	7	5	25
0, 0, 14, 14	7	7	49
2, 2, 2, 22	7	8.7	75.69

Statistika Deskriptif

Sifat Ragam dan Jangkauan

Sama seperti mean kedua ukuran tersebut sangat sensitif terhadap outlier. Tidak robust.

Statistika Deskriptif

Interquartile
Jarak antara kuartil 3 (Q3) dan kuartil 1 (Q1)

Q3 = ¾ th observasi
Q1 = ¼ th observasi

Data	Frequency
3	2
4	5
5	3
6	1
7	6
8	0
9	2
10	1
Total	20

Statistika Deskriptif

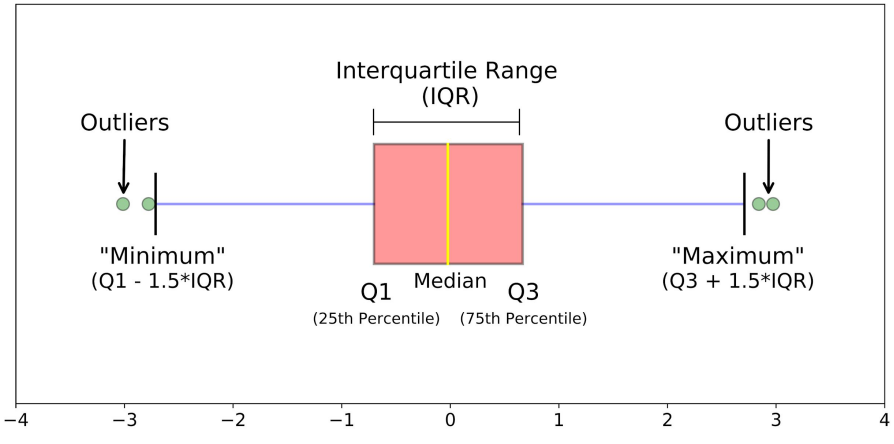
Interquartile
Jarak antara kuartil 3 (Q3) dan kuartil 1 (Q1)

Q3 = ¾ th observasi
Q1 = ¼ th observasi

Interquartile = Q3 - Q1 = 4-3 = 1

Data	Frequency
3	2
4	5
5	3
6	1
7	6
8	0
9	2
10	1
Total	20

Statistika Deskriptif



Sifat-sifat Interquartile

Interquartile adalah ukuran yang robust terhadap outlier.

Bahkan mampu mengidentifikasi dimana letak outlier berada.

Outline Pembelajaran

Rakamin Academy

Introduction to Statistics

- [x] Introduction to Statistics
- [x] Data Types
- [x] Descriptive Statistics
- [x] Measure of Central Tendency
- [x] Measure of Spread
- [ ] Hands-On

Rakamin Academy

Terima Kasih!