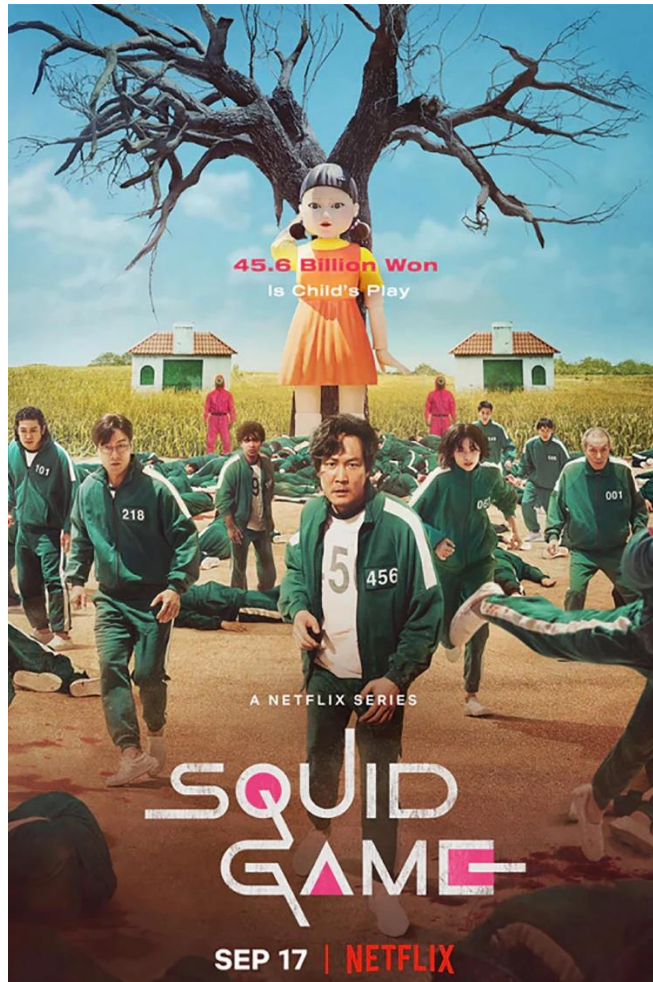




Statistika Deskriptif

Memahami data dengan lebih baik!



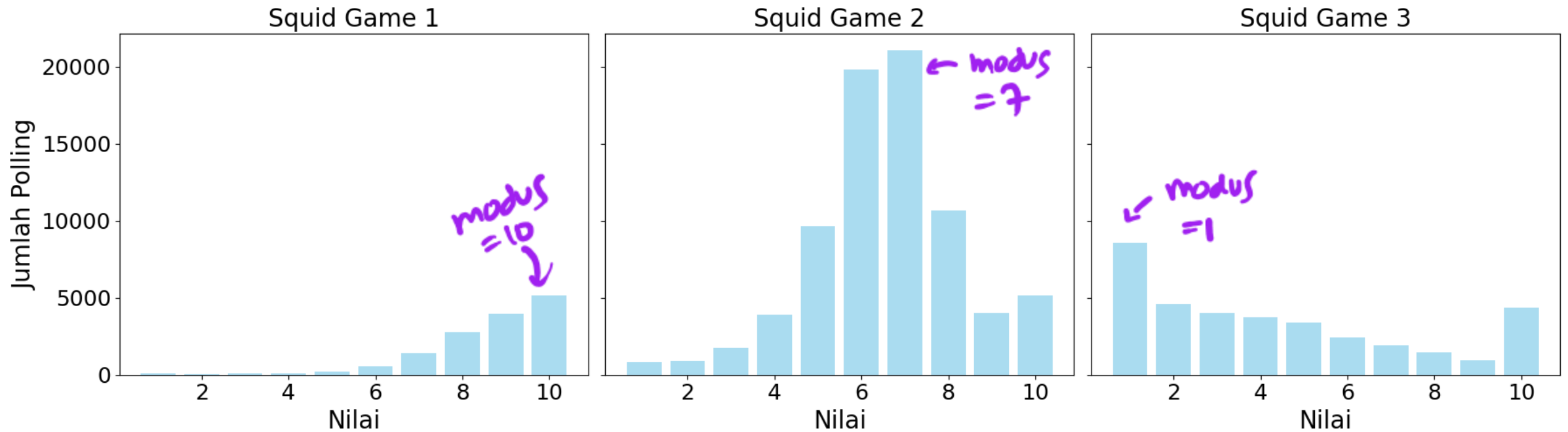
1.1 Variabilitas dan Visualisasi Data

Misalkan diberikan data sebagai berikut!

Nilai	Squid Game 1	Squid Game 2	Squid Game 3
10	5173	5169	4374
9	3936	4046	963
8	2792	10663	1471
7	1430	21076	1906
6	552	19837	2429
5	236	9669	3387
4	118	3933	3715
3	73	1720	3993
2	54	880	4612
1	124	859	8557

<https://colab.research.google.com/drive/1KDEGgYP8T-mnNuq8spRJVFJ3v9zU1EAS?usp=sharing>

Histogram untuk Masing-masing Film



1.2 Statistik atau Statistika?

- Statistik (*statistic*) adalah suatu besaran kuantitatif yang mewakili suatu populasi/dataset untuk memberikan gambaran secara ringkas mengenai populasi tersebut.
- Statistika (*statistics*) adalah ilmu yang mempelajari tentang Statistik.

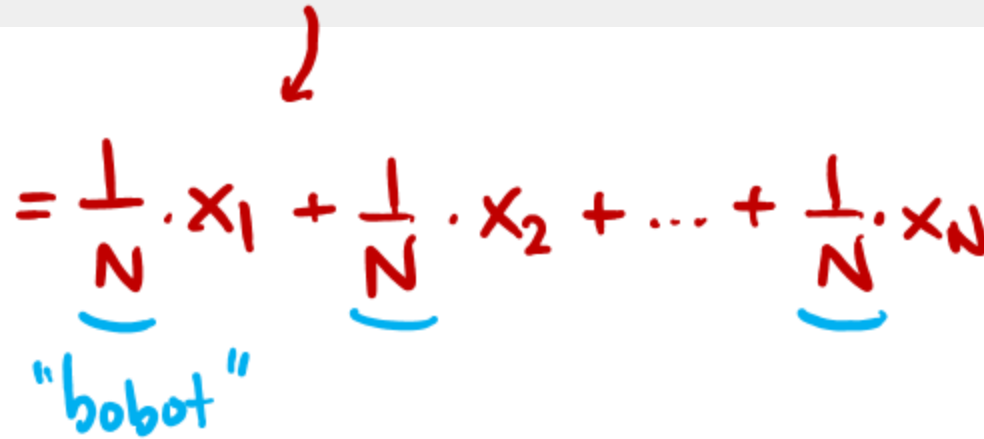


1.3 Ukuran Pusat Data

Definisi. Diberikan suatu data berukuran N dan secara terurut naik x_1, x_2, \dots, x_N .

(a) Rata-rata dari data tersebut didefinisikan sebagai

$$\bar{x} = \frac{1}{N}(x_1 + x_2 + \dots + x_N) = \frac{1}{N} \sum_{i=1}^N x_i.$$



Handwritten expansion of the average formula:

$$= \underbrace{\frac{1}{N}}_{\text{"bobot"}} \cdot x_1 + \underbrace{\frac{1}{N}} \cdot x_2 + \dots + \underbrace{\frac{1}{N}} \cdot x_N$$

1.3 Ukuran Pusat Data

Definisi. Diberikan suatu data berukuran N dan secara terurut naik x_1, x_2, \dots, x_N .

(b) Median dari data tersebut didefinisikan sebagai nilai yang berada di 'tengah' data, yakni

$$M_e = \begin{cases} x_{\frac{N+1}{2}}, & N \text{ ganjil,} \\ \frac{1}{2}(x_{\frac{N}{2}} + x_{\frac{N}{2}+1}), & N \text{ genap.} \end{cases}$$

$N=3$ - ganjil

x_1, x_2, x_3

median: $x_{\frac{3+1}{2}} = 2 //$

$N=4$ - genap

x_1, x_2, x_3, x_4

median: $\frac{1}{2}(x_{\frac{4}{2}} + x_{\frac{4}{2}+1})$
 $= \frac{1}{2}(x_2 + x_3)$

1.3 Ukuran Pusat Data

Definisi. Diberikan suatu data berukuran N dan secara terurut naik x_1, x_2, \dots, x_N .

(c) Modus dari data tersebut didefinisikan sebagai nilai yang paling sering muncul.

frekuensi paling besar.

Bagaimana cara mencari rata-rata, median, dan modus dari tabel film di atas?

⇒ Tabel Statistik:

	Statistik	Squid Game 1	Squid Game 2	Squid Game 3
0	Mean	8.621963	6.574243	4.339424
1	Median	9.000000	7.000000	4.000000
2	Mode	10.000000	7.000000	1.000000

→ rekomendasi

} ukuran pusat

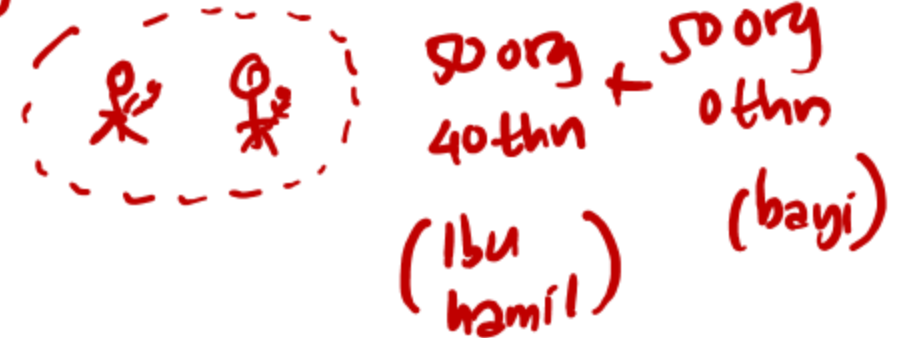
cukup??? ~~tidak!~~

"Acara ini akan dihadiri oleh 100 orang dengan usia rata-rata 20 tahun."

Ekspektasi



realita



1.4 Ukuran Variasi Data \bar{x}

Rata-rata, orang memberi nilai 8,6 untuk film Squid Game 1. Artinya, pada umumnya orang memberi skor di sekitar angka tersebut. Seberapa mungkin ada orang yang memberikan penilaian yang 'abnormal', yakni memberi skor yang cukup 'jauh' dari 8,6? Di sinilah, kita memerlukan ukuran variasi atau penyebaran. Misalkan Alisa memberi skor $x_A = 7$ dan Bobi memberi skor $x_B = 9$.

- Hitung simpangan kuadrat $(x_A - \bar{x})^2$. $\rightarrow (7 - 8,6)^2 = 2,56$
- Hitung simpangan kuadrat $(x_B - \bar{x})^2$. $\rightarrow (9 - 8,6)^2 = 0,16$
- Manakah yang memberi skor yang lebih 'normal' relatif terhadap keseluruhan data?



Definisi. Diberikan suatu data berukuran N dan secara terurut naik x_1, x_2, \dots, x_N .

(a) Variansi dari data tersebut didefinisikan sebagai

/Ragam

$$S^2 = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2.$$

rata-rata dari
simpangan kuadrat

(b) Standar Deviasi dari data tersebut didefinisikan sebagai

/simpangan baku

$$S = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}.$$

Statistik ukuran variansi menyatakan tingkat keberagaman dari sebuah data.

Dalam mata kuliah MATH1042, diketahui rata-rata nilai akhir mahasiswanya adalah 65. Robby memiliki nilai akhir sebesar 71. Berapa besar 'simpangan' nilai Robby dari nilai rata-rata?

(a) 6

(b) -6

~~(c) 36~~

(d) -36

(e) 71

'simpangan' $\rightarrow (x_R - \bar{x})^2 = (71 - 65)^2 = 6^2 = 36.$
 $\rightarrow (x_R - \bar{x}) = (71 - 65) = 6.$

Mengapa kita tidak mendefinisikan variansi menggunakan $\sum_{i=1}^N (x_i - \bar{x})$?

$$\begin{aligned} \sum_{i=1}^N x_i - \bar{x} &= \sum_{i=1}^N x_i - \left\{ \sum_{i=1}^N \bar{x} \right\}^{N\bar{x}} = \sum_{i=1}^N x_i - N\bar{x} = \sum_{i=1}^N x_i - N \left[\frac{\sum_{i=1}^N x_i}{N} \right] \\ &= \sum_{i=1}^N x_i - \sum_{i=1}^N x_i = 0 \end{aligned}$$

konstanta

utk data apapun, variansi = 0 (Absurd!)

Apakah bisa pakai $|x_i - \bar{x}|$? Ya!

Formula

>

$$\frac{1}{n} \sum_{i=1}^n |x_i - m(X)|$$

$m(X)$ = average value of the data set

n = number of data values

x_i = data values in the set

(simpangan rata²)

Berikut ini merupakan statistik dari nilai akhir mata kuliah MATH1061 dan MATH1062. Diketahui semua mahasiswa yang mengambil mata kuliah MATH1061 juga mengambil MATH1062.

Statistik	MATH1061	MATH1062
Rata-rata	80,43	73,07
Standar Deviasi	2,79	25,18

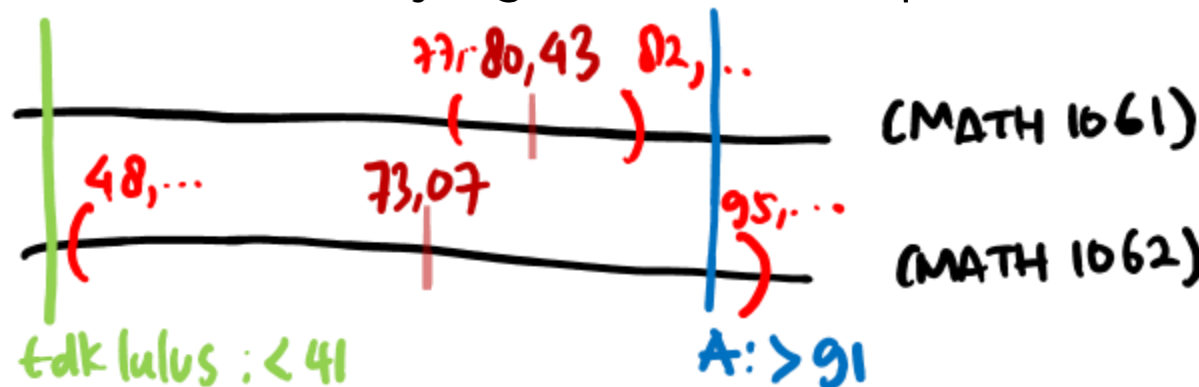
Catatan.

- Syarat mendapat nilai A, nilai akhir di atas 91.
- Syarat untuk kelulusan, nilai akhir di atas 41.

• Adakah yg dapat A di MATH1061?
 → tidak tahu → Ada jumlahnya sedikit sekali

Secara 'kasar', manakah kesimpulan yang tepat?

- (a) Banyak mahasiswa peroleh nilai A dari MATH1062 lebih banyak dari MATH1061. **Benar!**
- (b) Tidak ada mahasiswa yang tidak lulus baik pada MATH1062 maupun MATH1061. **Mungkin !!**



Nilai	Squid Game 2
10	5169
9	4046
8	10663
7	21076
6	19837
5	9669
4	3933
3	1720
2	880
1	859

pencilan }

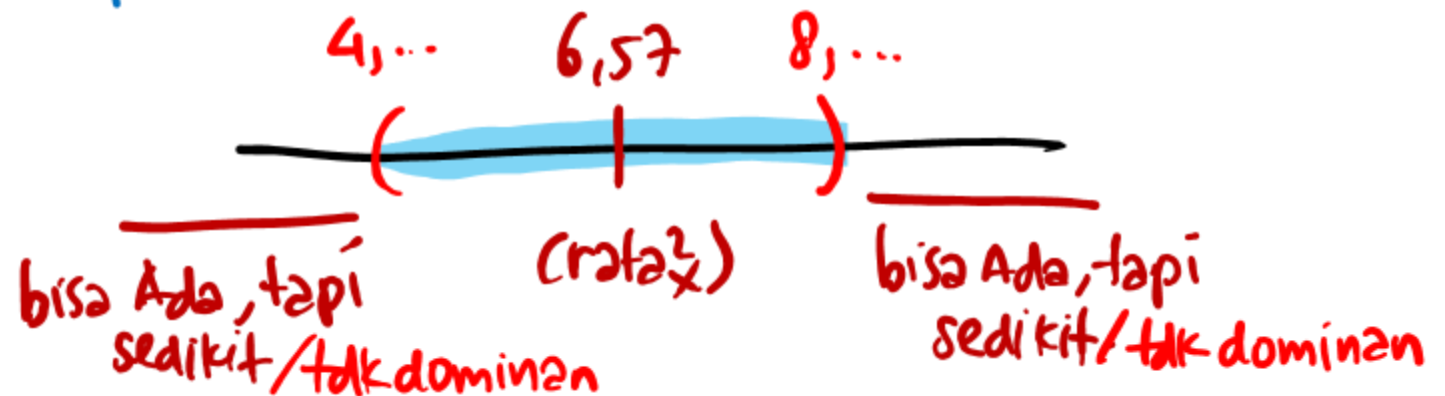
(lihat boxplot!)

} diluar simpangan baku, tapi belum termasuk pencilan!



Tabel Statistik:

	Statistik	Squid Game 2
0	Mean	6.574243
1	Median	7.000000
2	Mode	7.000000
3	Standard Dev	1.732251



Dua buah termometer diuji untuk mengukur benda yang diketahui secara absolut memiliki suhu sebesar 38° . Pengukuran dilakukan sama-sama sebanyak 100 kali untuk setiap termometer. Rata-rata hasil pengukuran kedua termometer sama-sama sebesar 38° . Standar deviasi dari pengukuran termometer pertama adalah 0,05 dan termometer kedua adalah 0,03. Manakah antara kedua termometer yang memiliki kualitas yang lebih bagus?

(a) Termometer pertama

(b) Termometer kedua

(c) Keduanya sama bagusnya

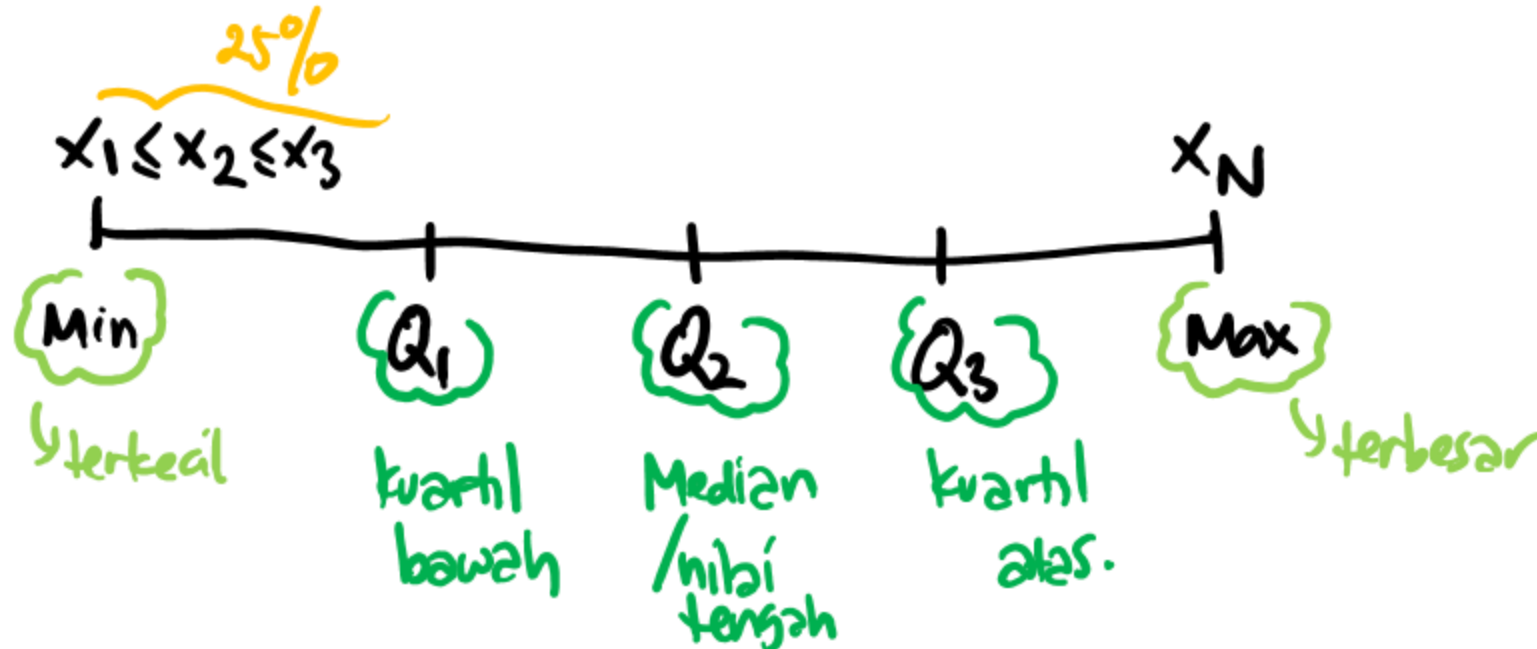
(d) Tidak cukup untuk menyimpulkan

1.5 Ukuran Lokasi dan Boxplot

← Penting!!!

Definisi. Misalkan suatu data berukuran N dan secara terurut naik x_1, x_2, \dots, x_N .

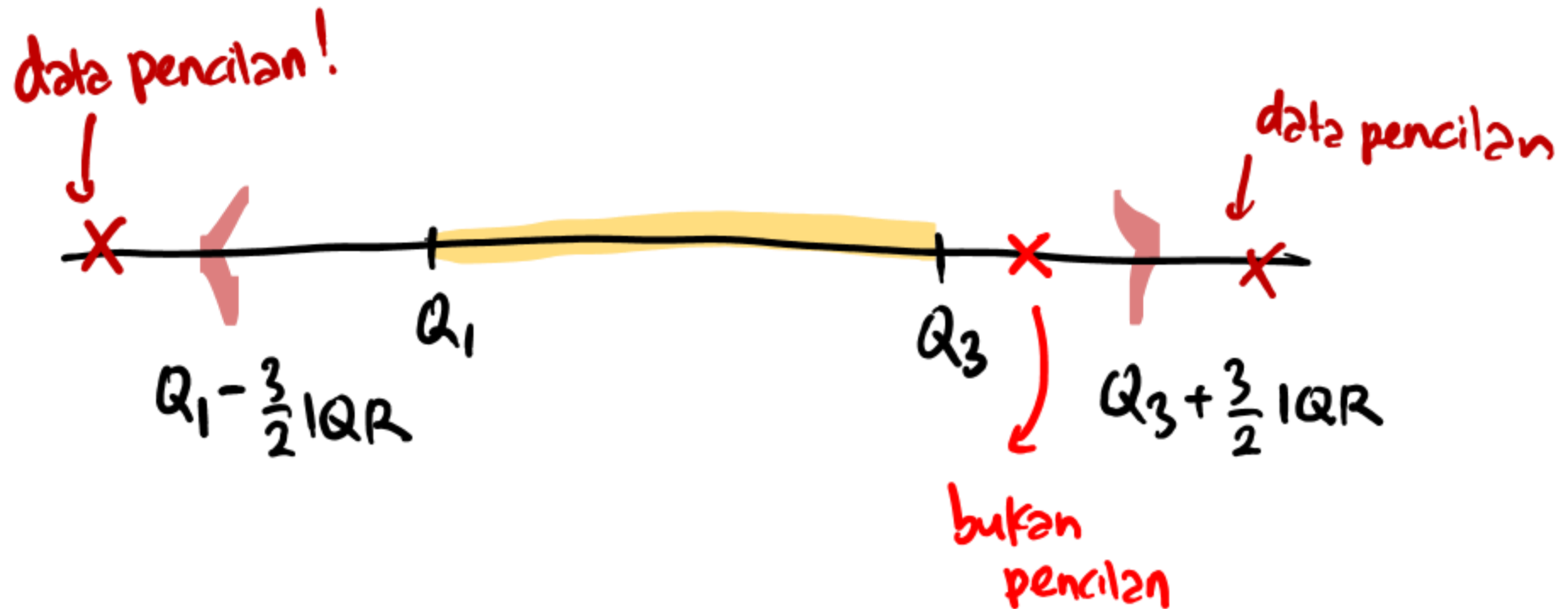
- Kuartil 1 dinotasikan sebagai Q_1 adalah nilai data ke 25% dari data terurut naik.
- Kuartil 2 dinotasikan sebagai Q_2 adalah nilai data ke 50% dari data terurut naik.
- Kuartil 3 dinotasikan sebagai Q_3 adalah nilai data ke 75% dari data terurut naik.



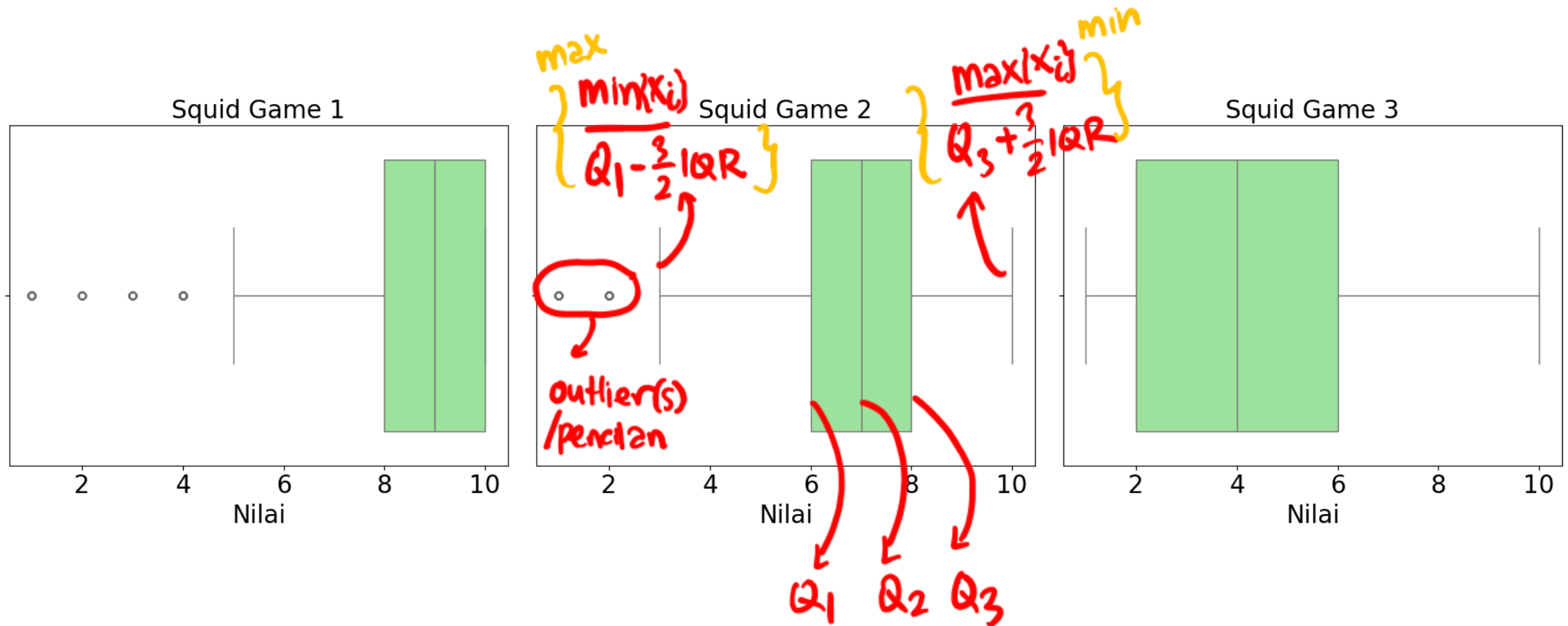
Definisi. Suatu data x_i disebut pencilan/outliers jika

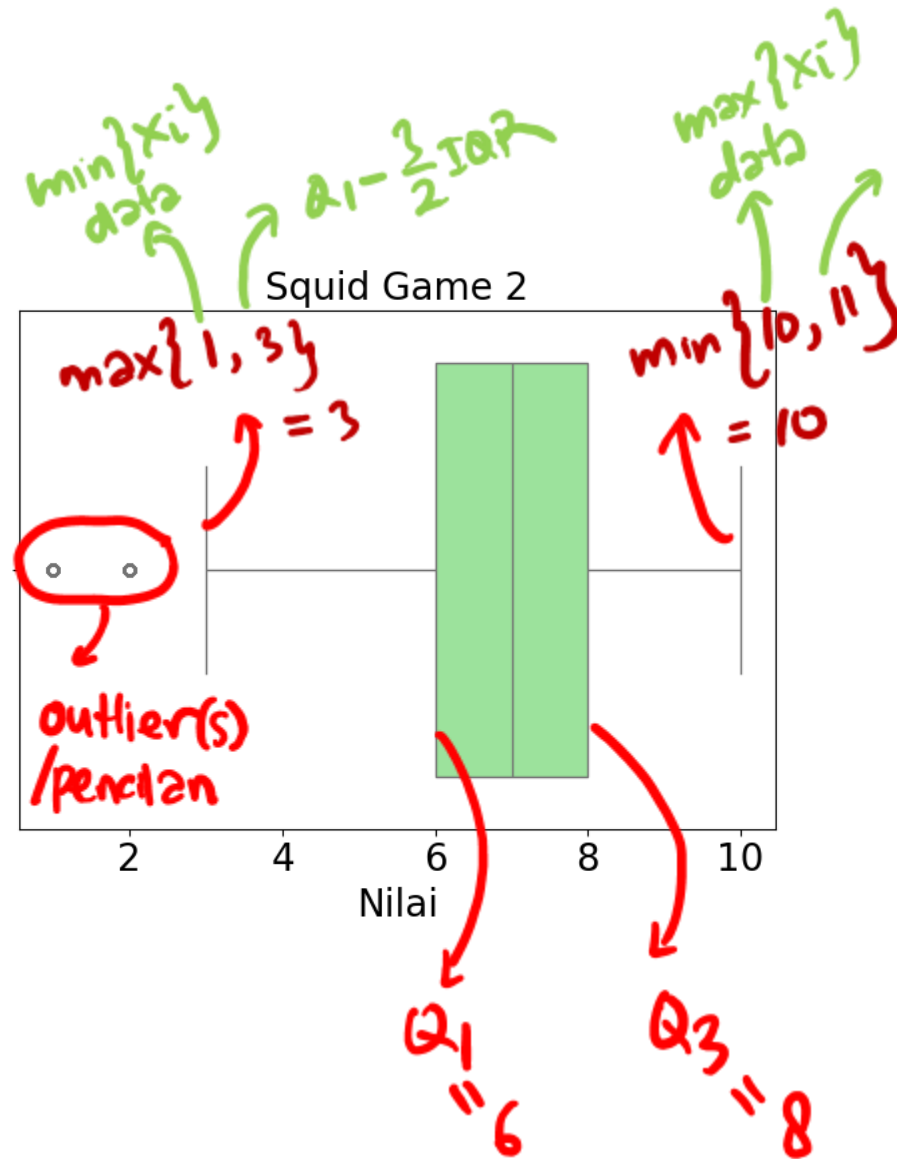
$$x_i < Q_1 - \frac{3}{2} \cdot IQR \quad \text{atau} \quad x_i > Q_3 + \frac{3}{2} \cdot IQR,$$

dengan $IQR = Q_3 - Q_1$ yang disebut sebagai nilai Jangkauan Antar kuartil.



Untuk memvisualisasikan **ukuran lokasi** dalam data, kita dapat menggunakan Bokplot.





Maka, $IQR = Q_3 - Q_1 = 8 - 6 = 2$.

Batas toleransi pencilan:

