Kursintroduktion

- Upplägg
  - Föreläsningar
  - Python-demos (Notebooks)
  - Räkneövningar
- Examination
  - Inlämningsuppgift (Python)
  - Tenta (Beräkning för hand)
- Kurs-PM

Kapitel 1

Introduktion till Statistisk Analys

Introduktion till Statistisk Analys

- Population och Stickprov
- Centraltendens
- Spridningsmått och kvartiler
- Datatyper

Population

Populationsdata beskriver *alla* värden i en data mängd:

- Alla invånare i Sverige
- Alla anställda på ett visst företag
- Alla besökare till en restaurang

Populationsdata är mycket sällan tillgängligt för statistiker.

- Urvalet för stort, eller oändligt
- Urvalet är destruktivt

Det totala antalet värden i en datamängd benämns *N*

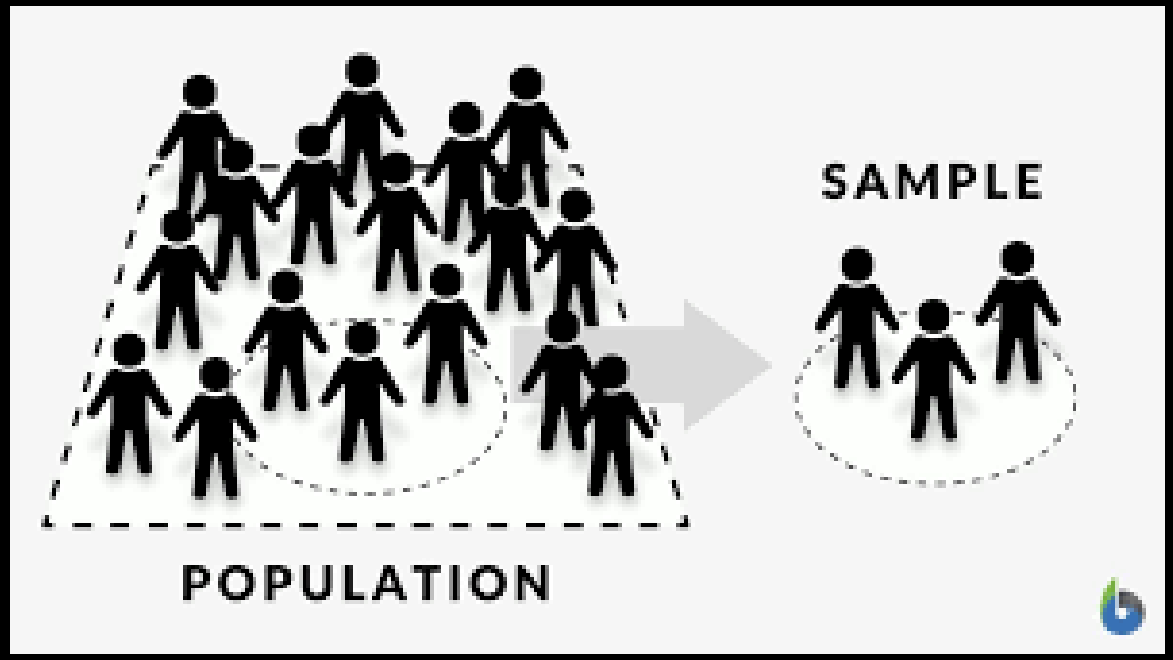
Stickprov (sample)

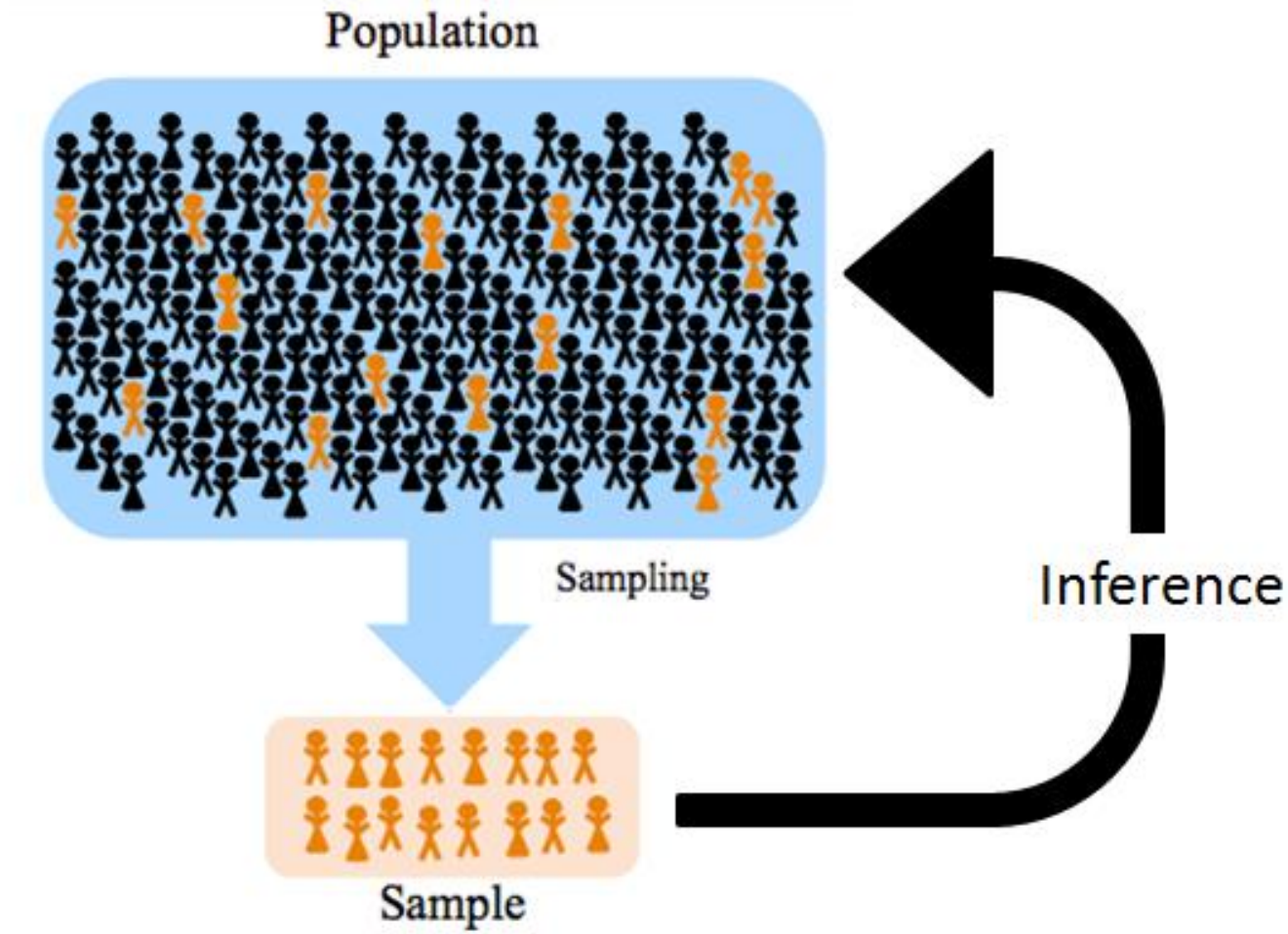
Ett stickprov är en mindre mängd observationer av en population.

- Några invånare i Sverige
- Ett urval av de anställda i ett visst företag
- En delmängd av samtliga besökare i en viss restaurang

De flesta beräkningar och uppskattningar i den här kursen kommer utgå från stickprov.

Antal värden i stickprov benämns *n*


SAMPLE
POPULATION


Population
Sampling
Sample
Inference

Population	Sample
• A population is the entire collection of subjects affected by your research question.	• A sample is a subset of the population you study.
• Measurements taken from a whole population are called parameters.	• Measurements taken from a sample are called statistics.
• Data for an entire population is often very difficult or impossible to collect.	• When population data is unavailable, we use sample data to make inferences about the population.
• If you do have data for a whole population, your parameters will be “true” measures of some population characteristic.	• Sample data yield statistics, which can be used to estimate population parameters. These estimates will always involve some margin of error due to sampling bias and other errors.

Anta att du är chef i en matbutik och vill mäta den generella kundnöjdheten. Du bestämmer dig därför för att ställa dig utanför butiken en tidig lördagsmorgon - precis vid öppningstid, och frågar därefter de första 10 kunderna om det är nöjda eller inte.

Av dessa är det 8 stycken som svarar, vad du uppfattar som, att de är nöjda.

Kan du dra slutsatsen att 80% av dina kunder generellt är nöjda med din butik?

Stickprov

Oberoende Slumpmässigt Urval OSU (eng. Independent Random Sample)

Ett OSU är ett urval från en population där alla datapunkter väljs slumpmässigt, med samma sannolikhet.

I praktiken inte alltid enkelt att genomföra, men en bra teoretisk utgångspunkt.

Centraltendens

Vad är ”centrum” i en datamängd?

- Typvärde
- Median
- Medelvärde (aritmetiskt medelvärde)

Typvärde

Typvärdet (eng. Mode) är det värde som förekommer oftast i en datamängd

Exempel - Bilar på en parkeringsplats har följande färger:

[Röd, Svart, Svart, Vit, Blå, Vit, Svart, Röd]

Typvärdet för färg på bilar är:

Svart (3 ggr)

Median

Medianen är det värde som hamnar i mitten av datamängden då den ordnats i storleksordning

Exempel: [3, 6, 4, 8, 4, 1, 3]

Sorterade värden:

[1, 3, 3, 4, 4, 6 ,8]

Medianvärdet är:

[1, 3, 3, 4, 4, 6 ,8]

Median

Medianen är det värde som hamnar i mitten av datamängden då den ordnats i storleksordning

Om antalet datapunkter är jämnt, är medianen mitt mellan mittenvärdena.

[3, 6, 4, 8, 4, 1, 3, 2] → [1, 2, 3, 3, 4, 4, 6 ,8] → (3 + 4) / 2 = 3.5

Medelvärde

Summan av alla värden, dividerat med antal värden.

För **populationer** betecknas medelvärdet med μ.

N är totalt antal värden i populationen.

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i = \frac{x_1 + \cdots + x_N}{N}$$

Exempel

Populationsvärden [3, 6, 4, 8, 4, 1, 3] $\rightarrow \mu = \frac{3+6+4+8+4+1+3}{7} \approx 4.14$

Medelvärde

Summan av alla värden, dividerat med antal värden:

För **stickprov** betecknas medelvärdet med $\bar{x}$.

n är totalt antal värden i stickprovet.

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n}$$

Populationsvärden [3, 6, 8, 4, 1, 3]

Stickprov [6, 8, 4] $\rightarrow \quad \bar{x} = \frac{6+8+4}{3} = 6$

Spridningsmått

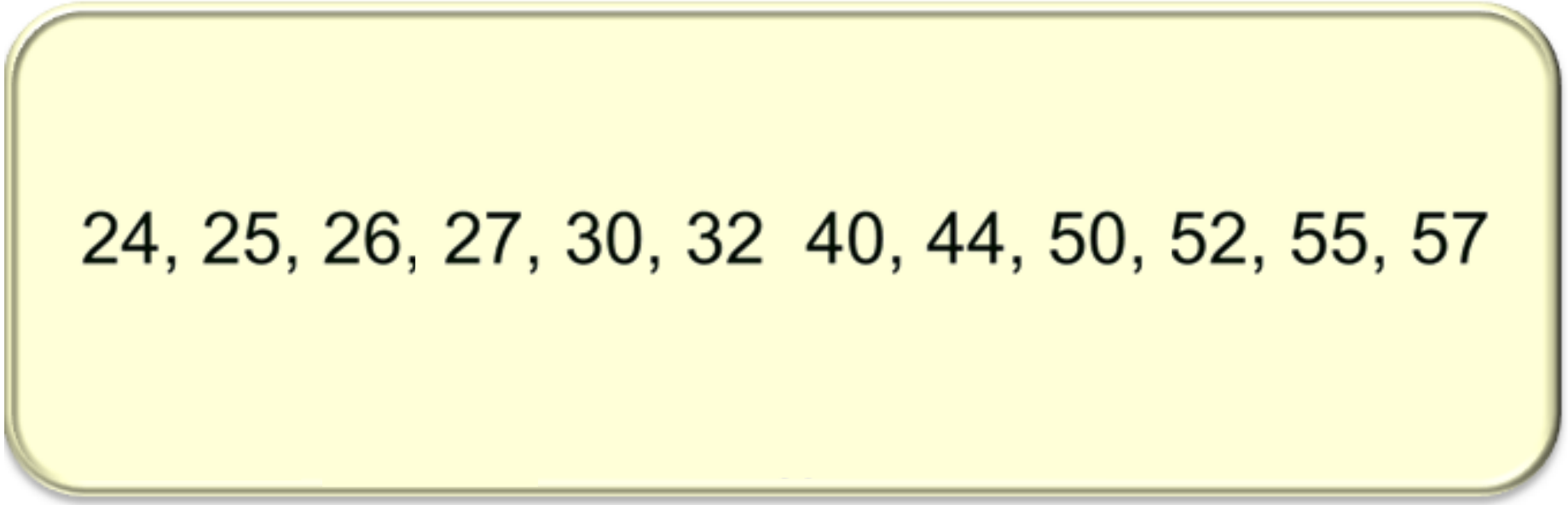
- Variationsvidd
- Kvartilavstånd

- Varians
- Standardavvikelse

- Medelabsolutavvikelse

Variationsvidd

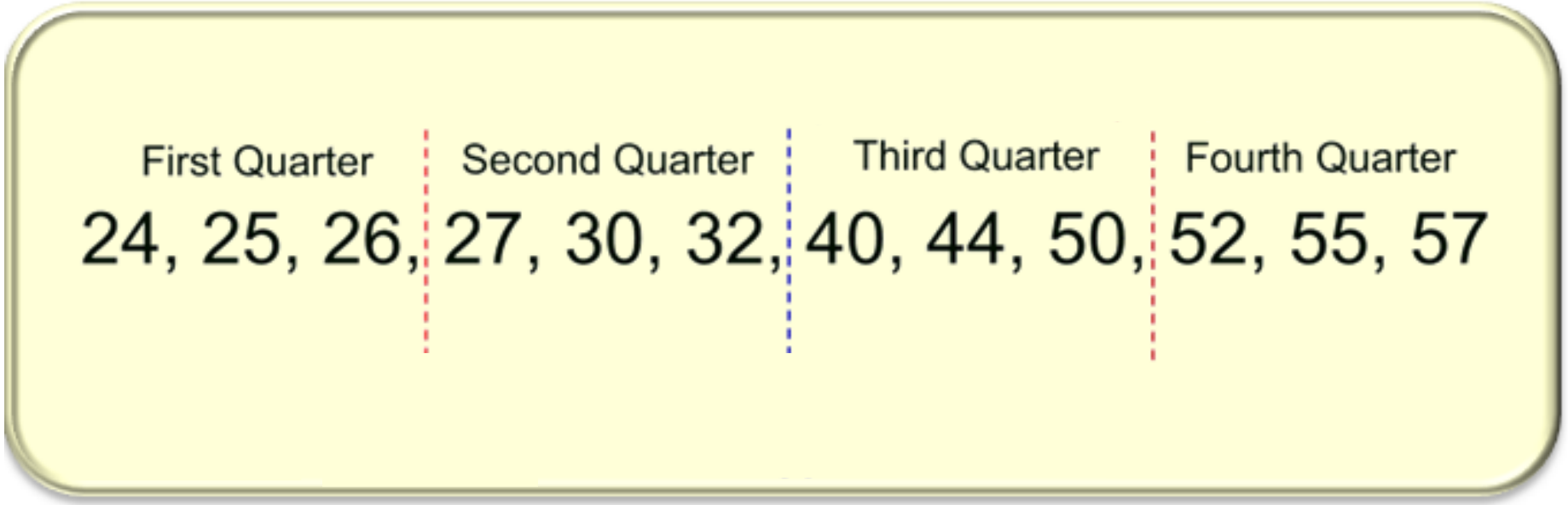
Helt enkelt avståndet mellan det högsta och det lägsta värdet



Variationsvidd = 57 – 24 = **23**

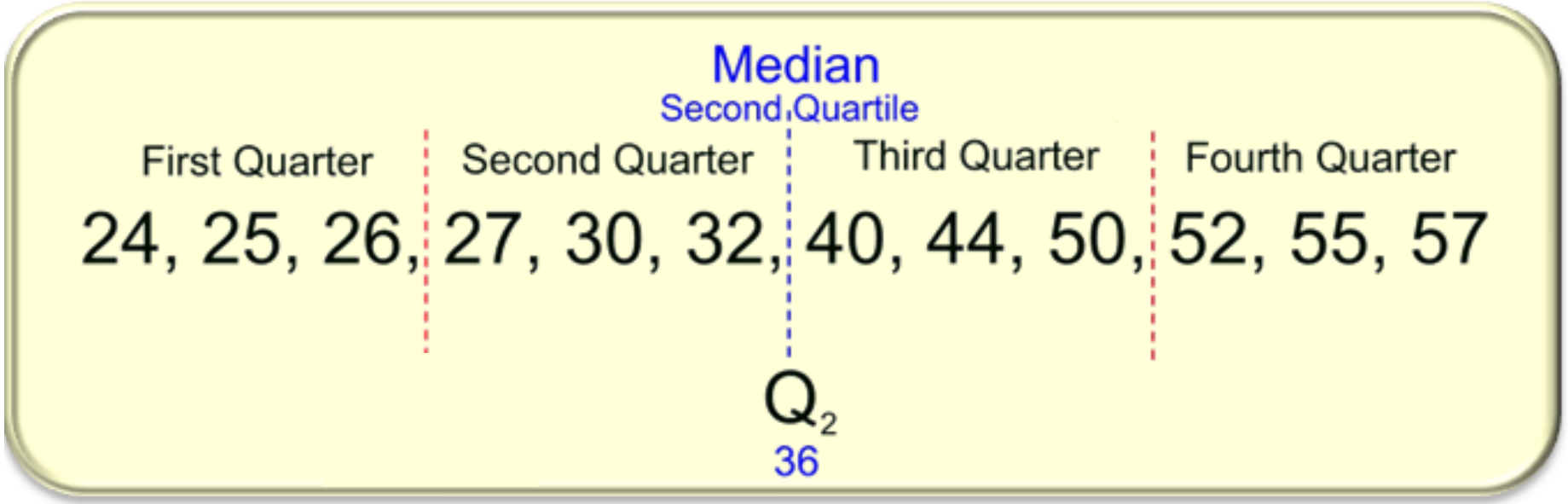
Kvartiler

Kvartiler används för att dela in data i fyra lika stora delar, s.k. kvarter.



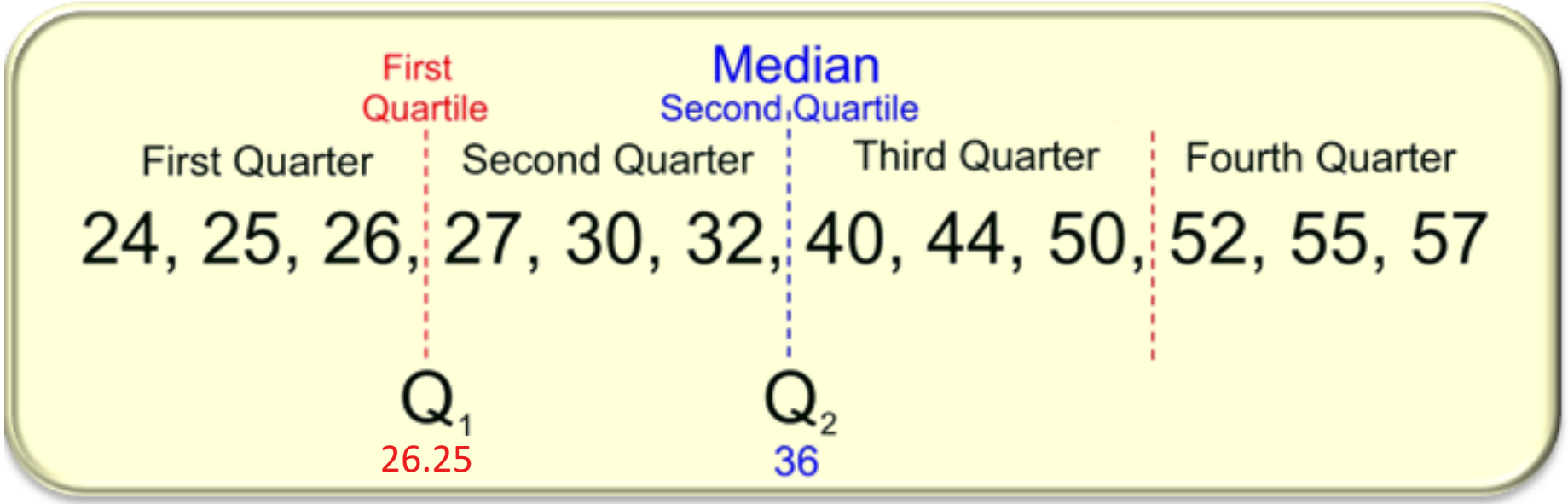
Kvartiler

Den andra kvartilen *är* medianen.



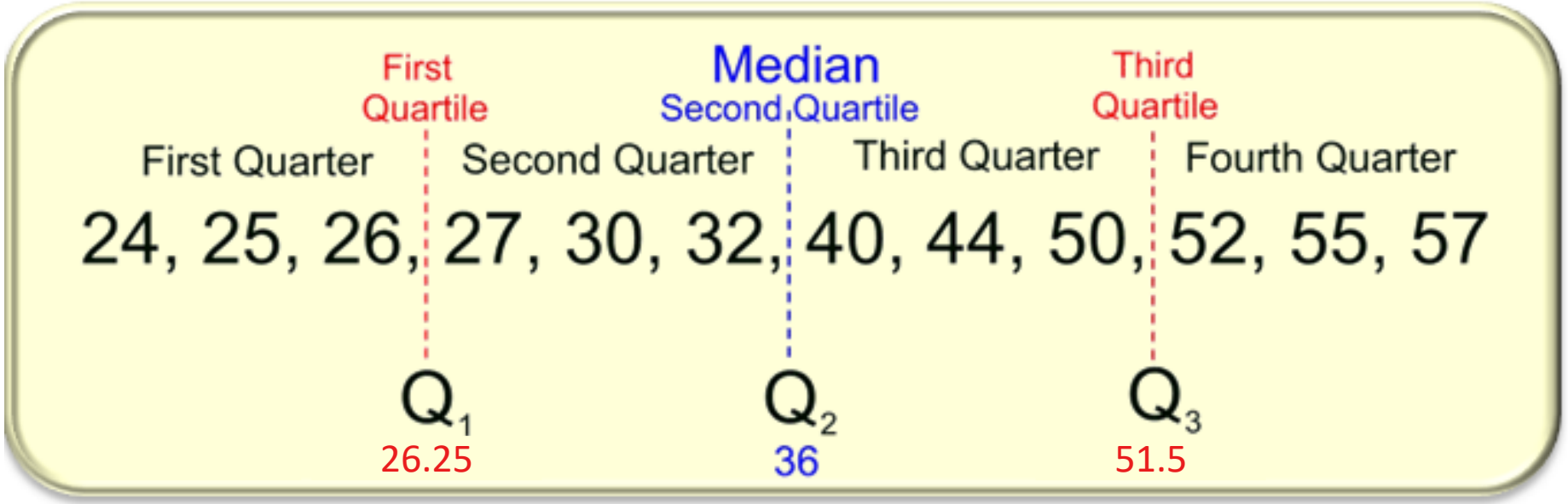
Kvartiler

Den **första** kvartilen är ’medianen’ av datamängden till vänster om medianen.



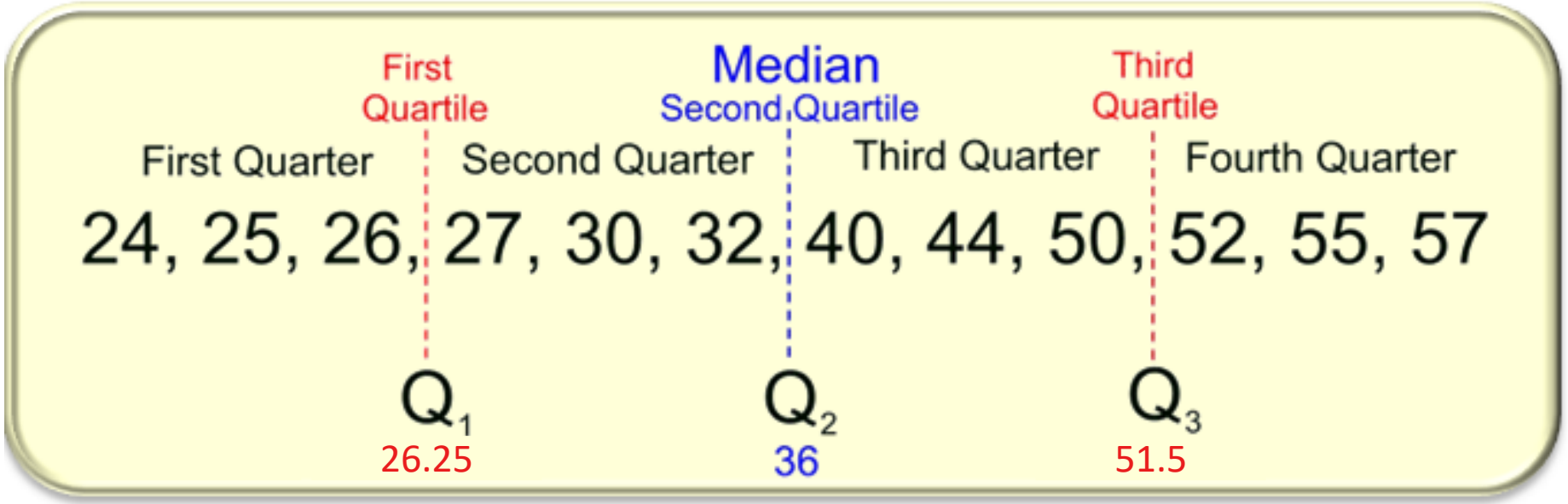
Kvartiler

Den **tredje** kvartilen är ’medianen’ av datamängden till höger om medianen.

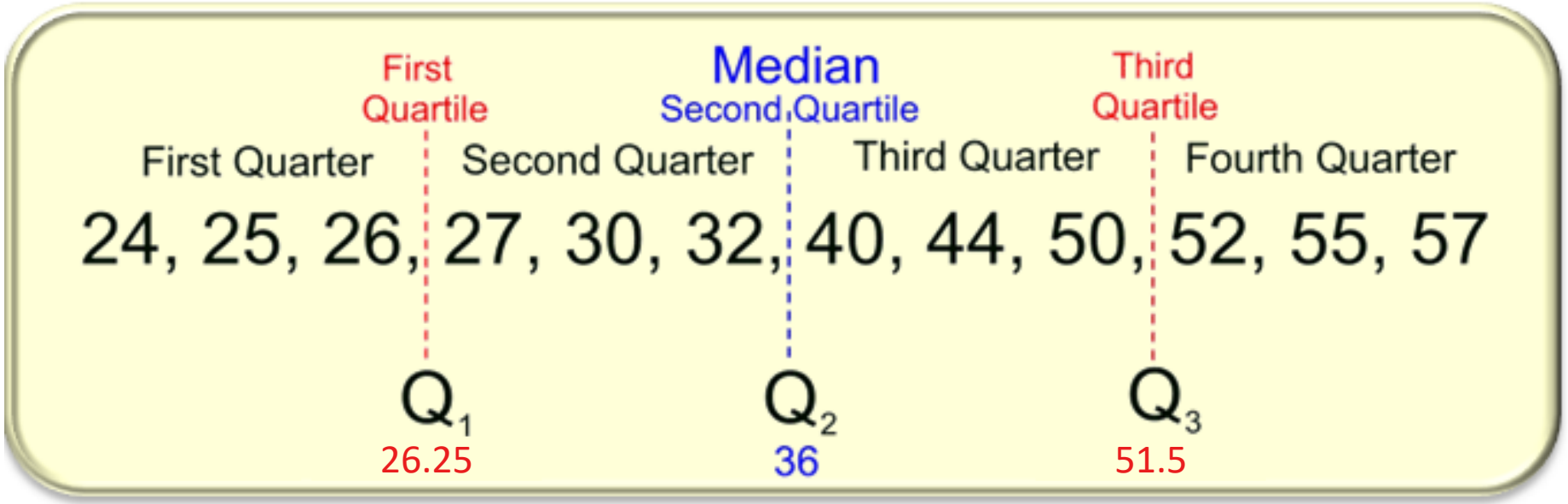


Kvartilavstånd

Kvartilavståndet (eng. Interquartile range, IQR) är avståndet mellan den första och tredje kvartilen.



$$Q_3 - Q_1 = 51.5 - 26.25 = 25.25$$



Anta n st datapunkter. I detta fall n = 12

Position för Q_2 (Median): $\frac{2(n+1)}{4} = \frac{2 \cdot 13}{4} = 6.5$

Medianen blir därför mitt-emellan siffrorna på position 6 och 7!

Position för Q1 : $\frac{(n+1)}{4} = \frac{13}{4} = 3.25$

Q_1 hamnar alltså **inte** exakt mittemellan siffrorna 26 och 27, utan bör vara närmare 26.

Resultatet blir att $\mathbf{Q_1}$ = 0.75 · 26 + 0.25 · 27 = **26.25**.

Position för Q_3 : $\frac{3(n+1)}{4} = \frac{3 \cdot 13}{4} = 9.75$

Q_3 hamnar alltså **inte** exakt mittemellan siffrorna 50 och 52, utan bör vara närmare 52.

Resultatet blir att $\mathbf{Q_3}$ = 0.25 · 50 + 0.75 · 52 = **51.5**.

Exempel

Antal sålda bilar per dag på Stures Bil AB de senaste 10 dagarna är 5, 5, 8, 4, 5, 4, 3, 3, 6, 11.

Beräkna variationsbredden, Q_1, Q_2, Q_3 samt kvartilavståndet.

Antal sålda bilar i sorterad ordning: 3, 3, 4, 4, 5, 5, 5, 6, 8, 11.

Antal datapunkter *n* = 10.

Variationsbredd = 11 – 3 = **8**

Position för Q_2 (Median): $\frac{2(n+1)}{4} = \frac{2 \cdot 11}{4} = 5.5$

Q_2 = **5**

Position för Q_1: $\frac{(n+1)}{4} = \frac{11}{4} = 2.75$

$Q_1 = 0.25 \cdot 3 + 0.75 \cdot 4 =$ **3.75**

Position för Q_3: $\frac{3(n+1)}{4} = \frac{3 \cdot 11}{4} = 8.25$

$Q_3 = 0.75 \cdot 6 + 0.25 \cdot 8 =$ **6.5**

Kvartilavstånd = $Q_3 – Q_1$ = 6.5 – 3.75 = **2.75**

Varians för en population med N datapunkter

Gäller för populationsdata

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Standardavvikelse

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Standardavvikelse är mer lättbegripligt än varians, och kan tolkas som ’medelavvikelsen’ av alla datapunkter, från medelvärdet.

Anta att vi har följande **populationsdata**. Det är antal korrekta svar på tentamen i linjär algebra, för samtliga elever, respektive. Klassen har alltså 7 st elever totalt. Dvs, *N = 7*.

Antal korrekta svar per elev: [10, 12, 14, 3, 9, 12, 11]

Medelvärde?

$$\mu = \frac{10 + 12 + 14 + 3 + 9 + 12 + 11}{7} \approx 10.1$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2$$

Varians?

$$\sigma^2 = \frac{(10-10.1)^2+(12-10.1)^2+(14-10.1)^2+(3-10.1)^2+(9-10.1)^2+(12-10.1)^2+(11-10.1)^2}{7} \approx 4.69$$

Standardavvikelse?

$$\sigma = \sqrt{\sigma^2}$$

$$\sigma \approx 2.17$$

Varians för en stickprov med n datapunkter

Gäller för stickprov

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

Standardavvikelse

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

Standardavvikelse är mer lättbegripligt än varians, och kan tolkas som ’medelavvikelsen’ av alla datapunkter, från medelvärdet.

Population Standard Deviation	Sample Standard Deviation
$$\sigma = \sqrt{\frac{\sum(x_i - \mu)^2}{N}}$$	$$s = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$
Where: σ is the population standard deviation $\sqrt{}$ is the symbol for taking the square root $\sum$ is the symbol for summation, indicating that you should take the sum of everything that follows it x_i is a particular value in your data μ is the population mean $(x_i - \mu)$ is the distance between a particular value in your population data and the mean N is the population size	Where: s is the sample standard deviation $\sqrt{}$ is the symbol for taking the square root $\sum$ is the symbol for summation indicating, that you should take the sum of everything that follows it x_i is a particular value in your data $\bar{x}$ is the sample mean $(x_i - \bar{x})$ is the distance between a particular value in your sample data and the mean n is the sample size*

*When calculating the sample standard deviation, we divide by n-1 to get a closer approximation of the true population standard deviation,σ.

Medelabsolutavvikelse

Medelabsolutavvikelse (eng. Mean Absolute Error MAE) är medelvärdet av *absolutavvikelser* från medelvärdet.

$$MAD = \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}|$$

MAD är mer *robust* än standardavvikelse, och används ibland inom utveckling av prediktionsmodeller.

Datatyper

Data kategoriseras på olika sätt:

- Nominal
- Ordinal
- Intervall
- Kvot

Ordinal

Ordinaldata är sådan data om går att kategorisera, och där det även finns en inbördes rangordning mellan kategorierna.

Exempelvis:

Expertise - {Beginner, Intermediate, Expert}

Education level - {Primary, Secondary, Undergraduate, Graduate}

Income - {Low, Medium, High}

Nominal

Nominaldata är sådan data som går att kategorisera – men där det inte existerar någon inbördes rangordning.

Nationality - {Swedish, Norwegian, Danish, Finnish}

Favorite genre – {Horror, Comedy, Action, Thriller, Romance}

Interval

Intervaldata är sådan data som det går att mäta skillnaden mellan värden – men som **inte** har en meningsfull nollpunkt

Temperature (C) – {-273.15, .., -10, .., 0, .., 10, .., 1 000 000, ..}

Vad means med en meningsfull nollpunkt?

Exempelvis är skillnaden mellan 1 C och 16 C exakt 16x, men det betyder inte att 16 C är 16 gånger större än 1 C.

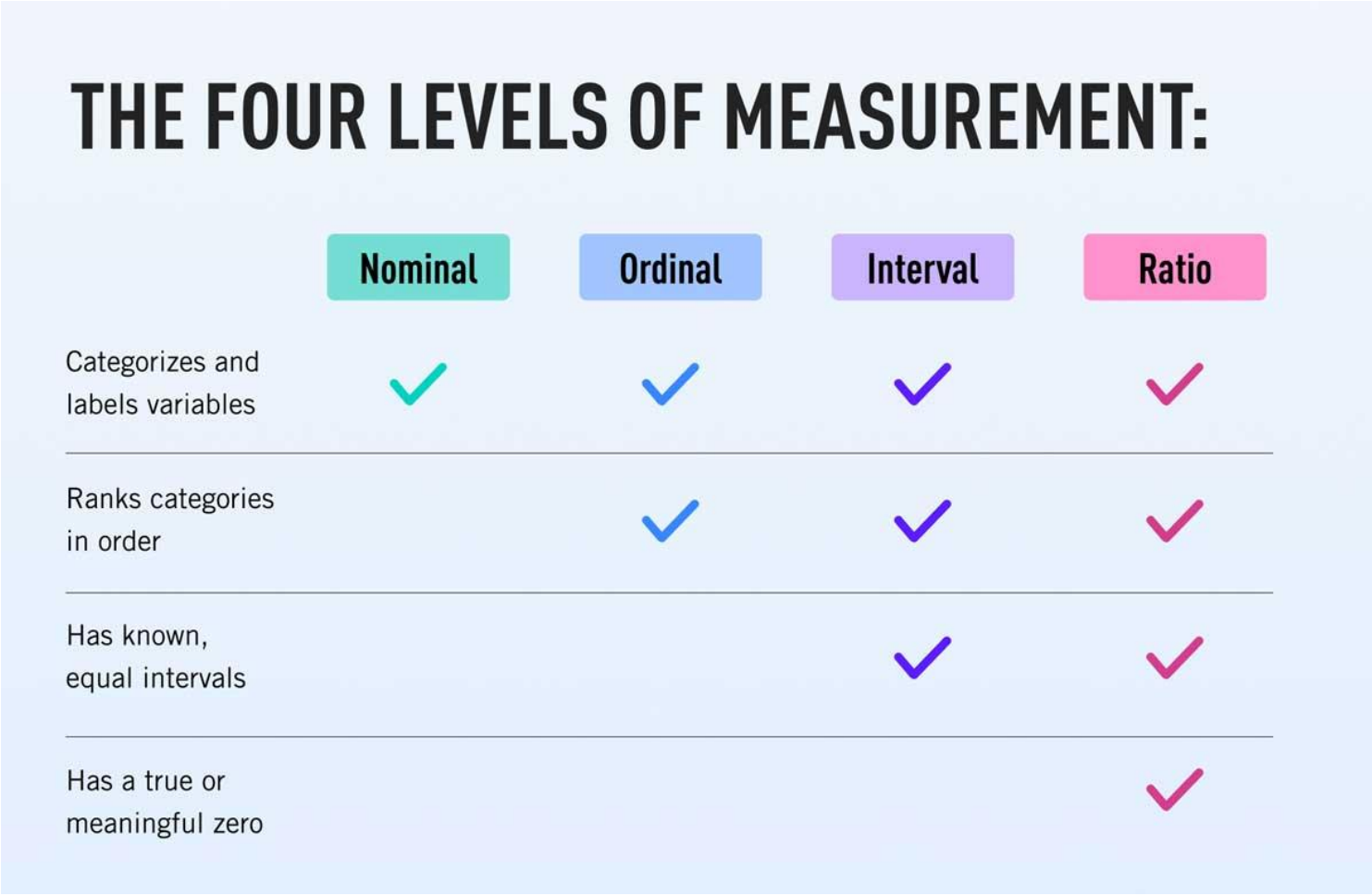
Kvot

Kvotdata är sådan data som det går att mäta skillnaden mellan värden – och som har en meningsfull nollpunkt

Födelsevikt (g) – {0, 1, 2, .., 1000, .., 2000, .., 3000, ..}

Topphastighet (km/h) – {0, 1, 2, 3, .., 100, .., 500, ..}

Datatyper


THE FOUR LEVELS OF MEASUREMENT:
Nominal
Ordinal
Interval
Ratio
Categorizes and labels variables
Ranks categories in order
Has known, equal intervals
Has a true or meaningful zero