Kapitel 1

Introduktion till Statistisk Analys

Introduktion till Statistisk Analys

- Kursintroduktion
  - Kurs-PM
  - Upplägg
  - Examination
- Datatyper
- Population och Stickprov
- Centraltendens
- Spridningsmått och kvartiler

Kurinstroduktion

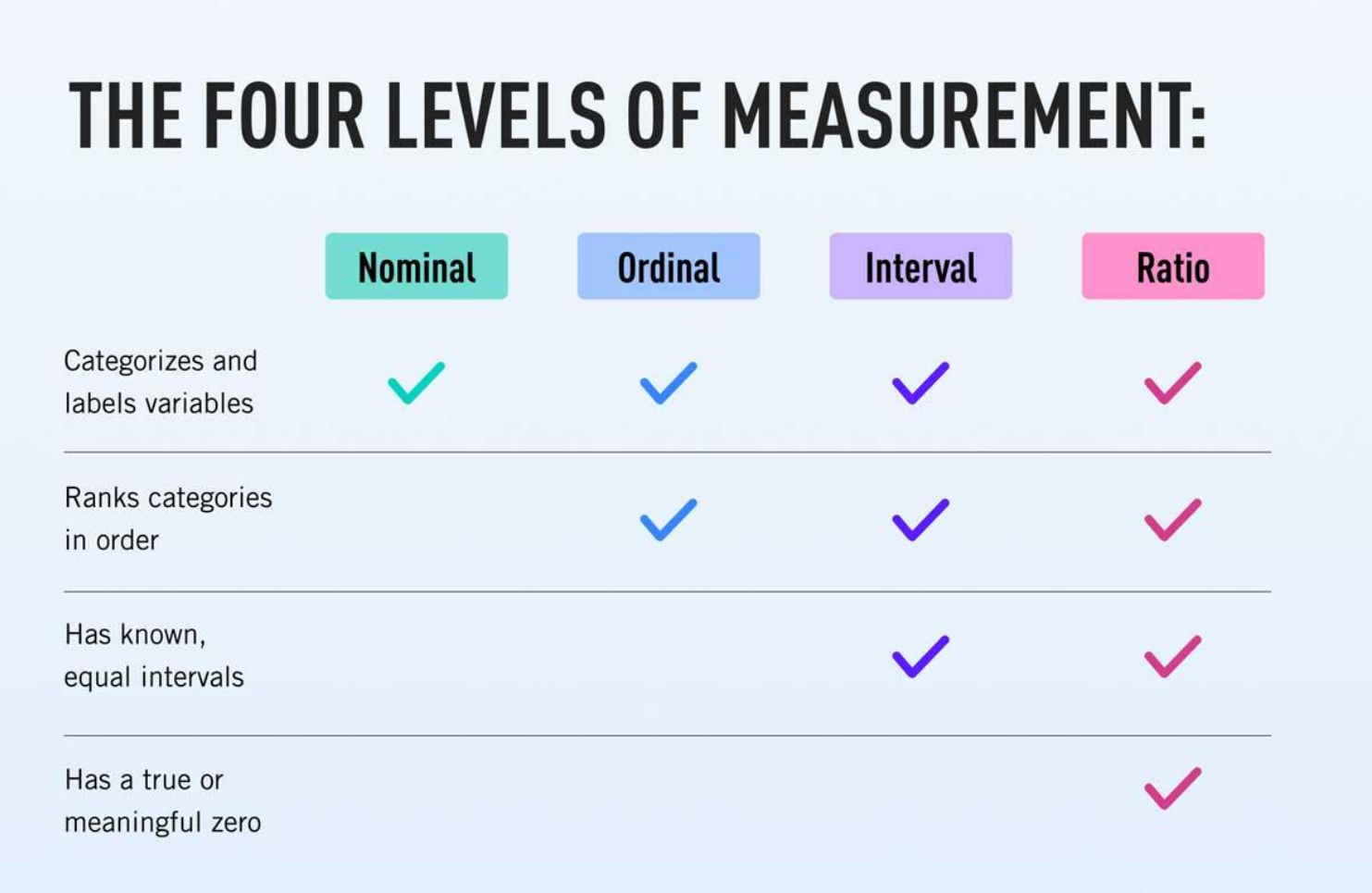
- Upplägg
  - Föreläsningar
  - Python-demos (Notebooks)
  - Räkneövningar
- Examination
  - Inlämningsuppgift (Python)
  - Tenta (Beräkning för hand)
- Kurs-PM

Datatyper

Data kategoriseras på olika sätt:

- Nominal
- Ordinal
- Intervall
- Kvot

- Diskret
- Kontinuerlig



https://careerfoundry.com/en/blog/data-analytics/what-is-interval-data/

Populationsdata

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 statistisker.

- Urvalet för stort, eller oändligt
- Urvalet är destruktivt

Stickprov

Ett stickprov (eng. sample), är en mindre mängd observationer av en population.

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

Att genomföra stickprov på ett korrekt sätt kan vara en hel kurs i sig självt, och beror på många faktorer.

- Statistisk styrka
- Ekonomi
- Datahantering
- Databehandling
- Etik

Stickprov

Oberoende slumpmässigt urval (OSU) (eng. Simple Random Sample, SRS)

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)

- Geometriskt medelvärde
- Harmoniskt medelvärde
- Interkvartilt 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}

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

Exempel: {3, 6, 4, 8, 4, 1, 3}

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

Formellt, aritmetiskt medelvärde (eng. arithmetic mean)

Summan av alla värden, dividerat med antal värden:

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

Exempel {3, 6, 4, 8, 4, 1, 3} → $\mu = \frac{3+6+4+8+4+1+3}{7} \approx 4.14$

Om medelvärdet är för ett stickprov betecknas det istället med $\bar{x}$.

Spridningsmått

- Varians
- Standardavvikelse

- Kvartilavstånd
- Medelabsolutavvikelse

Varians

Den normerade kvadrerade avvikelsen från medelvärdet:

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

För ett stickprov:

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

Skillnaden i nämnaren är relaterat till antalet *frihetsgrader* (eng. *degrees of freedom)*

Varians

Antalet frihetsgrader avgörs av hur många okända parametrar vi tar med oss in i beräkningen.

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

Eftersom medelvärdet μ är beräknat utifrån stickprovet korrigerar vi för detta genom att sätta antalet frihetsgrader till $n - 1$ istället för n.

Varians

En mindre beräkningsintensiv form för stickprovsvarians:

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

Standardavvikelse

Standardavvikelse (eng. standard deviation) är kvadratroten av variansen.

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

Standardavvikelse är mer lättbegripligt. Jämför enheterna för varians och standardavvikelse

Kvartilavstånd

Kvartiler är som medianer, fast uppdelat i fyra. Den andra kvartilen *är* medianen.

https://www.scribbr.com/statistics/quartiles-quantiles/

Kvartilavstånd

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

https://www.scribbr.com/statistics/quartiles-quantiles/

Kvartilavstånd

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

https://www.scribbr.com/statistics/quartiles-quantiles/

Kvartilavstånd

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

https://www.scribbr.com/statistics/quartiles-quantiles/

Medelabsolutavvikelse

Medelabsolutavvikelse (eng. Mean Absolute Deviation – MAD) 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.