Beskrivelse af stikprøver og populationer (fra kurset ST520)

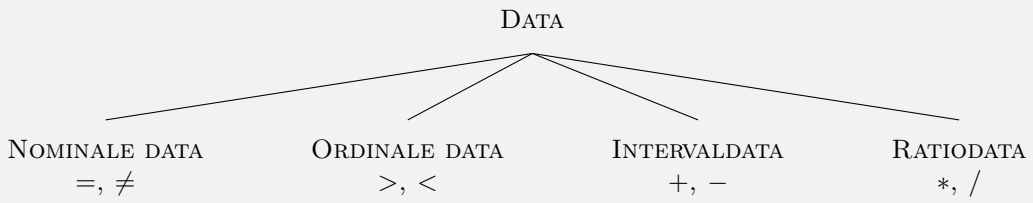
Christian Damsgaard Jørgensen

Forår 2017

Datatyper

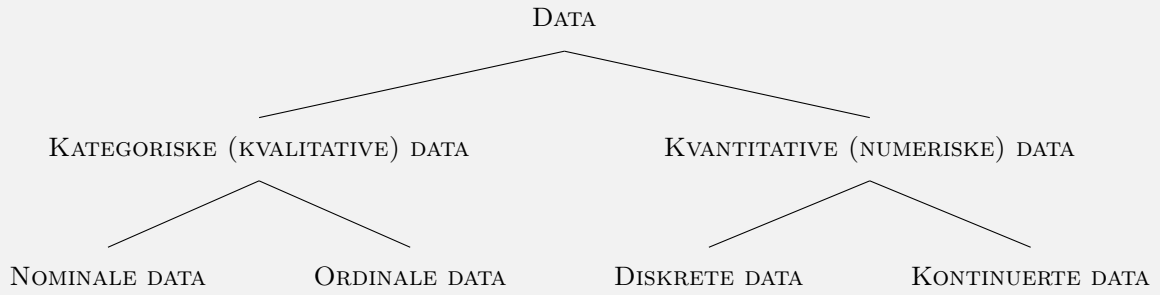
Den mest kendte klassifikation af data:

Stevens (1946):



Lærebogens klassifikation af data:

Ekstrøm og Sørensen (2014):

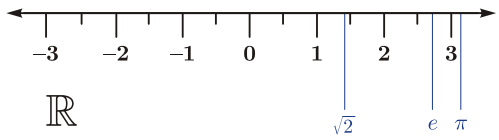


Nominale data: Når der ikke findes en naturlig rækkefølge af kategorierne.

Ordinale data: Når kategorierne kan ordnes.

Diskrete data: Numeriske datavariable med et endeligt eller tælleligt antal mulige værdier.

Kontinuerte data: Numeriske datavariable med alle tal i et interval på den reelle tallinje som mulige værdier.



Visualisering af kategoriske (kvalitative) data

Hyppighed

Antallet af forekomster af hver værdi i et datasæt.

Søjlediagram

Simpelt plot, der viser de mulige kategorier og hyppigheden af hver kategori.

Frekvens (eller relativ hyppighed)

Beregnes ved at dividere hyppigheden af en kategori med antallet af observationer i stikprøven:

$$\text{frekvens} = \frac{\text{hyppighed}}{n}$$

Segmenteret søjlediagram

Viser frekvenser af kategorierne i en stikprøve som én søjle med en samlet højde på 100%.

Visualisering af kvantitative (numeriske) data

Histogram

Grafisk opsummering af et datasæts fordeling svarende til et søjlediagram (med hyppighed på y-aksen).

En variant af histogrammet med frekvens fremfor hyppighed på y-aksen kaldes et frekvenshistogram.

Kontinuerte data kan grupperes i passende intervaller – hvorved man kan tælle antallet af observationer, som falder indenfor hvert interval. De resulterende intervaller og de tilhørende frekvenser giver variablens fordeling.

Scatterplot

Illustration af forholdet mellem to kvantitative variable, hvor datapunkter plottes i et koordinatsystem.

Opsummerende statistik

Lad $y_1, y_2, \ldots, y_n$ betegne de kvantitative observationer i en stikprøve af størrelse n fra en population. Vi kan sortere observationerne fra mindste til største værdi og benytter i det følgende notationen $y_{(1)}, y_{(2)}, \ldots, y_{(n)}$ til at repræsentere et ordnet observationssæt – hvor $y_{(1)}$ er den mindste værdi, $y_{(2)}$ er den næstmindste værdi osv.

Central tendens

Repræsenterer værdien af en typisk observation i et datasæt.

Dispersion (eller variabilitet)

Repræsenterer hvor meget observationerne i et datasæt afviger fra den centrale tendens.

Median

Det midterste tal eller gennemsnittet af de to midterste tal, når tallene er sorterede:

$$\text{median} = \begin{cases} y_{\left(\frac{n+1}{2}\right)} & \text{hvis } n \text{ er ulige,} \\ \frac{1}{2}\left[y_{(n/2)} + y_{(n/2+1)}\right] & \text{hvis } n \text{ er lige.} \end{cases}$$

50% af observationerne er mindre end medianen og 50% af observationerne er større end medianen.

Medianen er et mål for central tendens.

Variationsbredde

Forskellen mellem den største observation og den mindste observation:

$$\text{variationsbredde} = y_{(n)} - y_{(1)}$$

Variationsbredden er et mål for dispersion.

Første kvartil (eller nedre kvartil)

Betegnes med Q1. 25% af observationerne er mindre end Q1 og 75% af observationerne er større end Q1.

Tredje kvartil (eller øvre kvartil)

Betegnes med Q3. 75% af observationerne er mindre end Q3 og 25% af observationerne er større end Q3.

Kvartilafstand

Forskellen mellem den tredje kvartil og den første kvartil:

$$\text{IQR} = \text{Q3} - \text{Q1}$$

Kvartilafstanden er et mål for dispersion.

Fraktil

Den k'te fraktil opdeler et observationssæt, sådan at $k/100$ af observationerne er mindre end k.

Bemærk at medianen er et specialtilfælde af kvartiler, som igen er et specialtilfælde af fraktiler.

Boksplot (eller kassediagram)

Grafisk opsummering af et datasæt.

Her plottes den mindste observation, den første kvartil, medianen, den tredje kvartil og den største observation.

Ekstreme observationer

Som tommelfingerregel er en observation ekstrem, hvis den er mindre end $1{,}5 \cdot \text{IQR}$ under den første kvartil eller større end $1{,}5 \cdot \text{IQR}$ over den tredje kvartil, dvs. hvis den falder udenfor intervallet $[\text{Q1} - 1{,}5 \cdot \text{IQR}; \text{Q3} - 1{,}5 \cdot \text{IQR}]$.

Modificeret boksplot

Et boksplot, hvor ekstreme observationer plottes som individuelle punkter – og hvor minimum og maksimum er erstattet med de mindste og største observationer, som falder indenfor intervallet $[\text{Q1} - 1{,}5 \cdot \text{IQR}; \text{Q3} - 1{,}5 \cdot \text{IQR}]$.

Stikprøvemiddelværdi

Summen af alle observationer i stikprøven divideres med antallet af observationer i stikprøven:

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

Stikprøvemiddelværdien er et mål for central tendens.

Stikprøvestandardafvigelse

Løst fortalt måler stikprøvestandardafvigelsen den "gennemsnitlige" afvigelse fra stikprøvens middelværdi:

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

Bemærk at det ville have været en gennemsnitlig afvigelse, hvis vi i stedet havde divideret summen med n.

Stikprøvestandardafvigelsen er et mål for dispersion.

Stikprøvevarians

Stikprøvevariansen er den kvadrerede stikprøvestandardafvigelse:

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

Stikprøvevariansen er et mål for dispersion.

Litteratur

[1] Ekstrøm, Sørensen. *Introduction to Statistical Data Analysis for the Life Sciences.* CRC Press, 2014.

[2] Stevens. *On the Theory of Scales of Measurement.* Science (vol. 103, no. 2684, pp. 677-680), 1946.