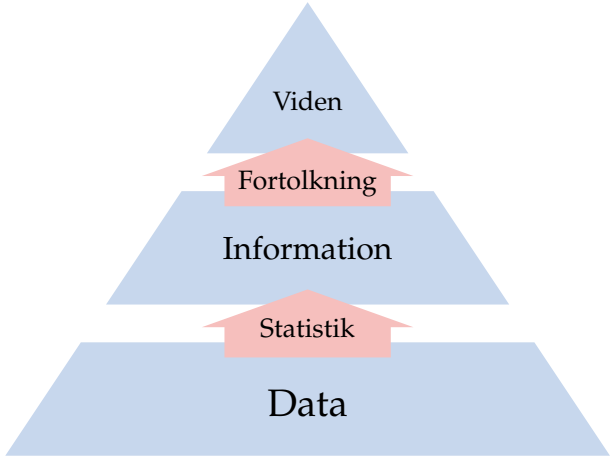


Datasæt

Statistik er en gren af matematik hvor man arbejder med *data*. Selve ordet *data*, der fra latin betyder *det som er givet*, kan for det meste sidestilles med ordet observationer.

Data indsamles typisk fordi man vil opnå en viden om det man indsamler data omkring. Men før man kan opnå viden fra de indsamlede data, skal det først analyseres med statistiske metoder. Resultatet af de statistiske undersøgelser er information og denne information skal fortolkes korrekt før man til sidst opnår den viden man er interesseret i.



Data forekommer på mange forskellige former alt efter hvilke observationer dataværdierne angiver. De forskellige typer af data kan inddeles i kategorier hvor den første inddeling består af hhv. kvalitativ data og kvantitativ data. *Kvalitativ data* består af beskrivelser som f.eks. hvordan en række af testpersoner beskriver smagen af en ny type lakrids. *Kvantitativ data* består af talværdier og kan opdeles i to yderligere kategorier: *optællinger* og *målinger*.

Et *datasæt* er essentielt et ordnet sæt X af et antal af n *dataværdier* $x_1, x_2, \dots, x_n$. Alle datasættets dataværdier er af samme *datatype*, hvilket betyder at der findes en mængde $\mathbb{X}$ som angiver hvilke dataværdier der kan forekomme i datasættet:

$$X = (x_1, x_2, \dots, x_n) \quad \text{hvor} \quad x_1, x_2, \dots, x_n \in \mathbb{X}$$

Betragter man for eksempel et datasæt, der f.eks. kunne fremkomme ved at undersøge ti husstande for hvor mange fjernsyn hver husstand har, kunne se således ud:

$$X = (2, 3, 3, 1, 4, 5, 4, 3, 2, 4)$$

Datasættets størrelse (dvs. antallet af dataværdier) er 10, og at alle dataværdierne tilhører datatypen af positive heltal.

Et andet eksempel på et datasæt, der f.eks. kunne fremkomme ved at undersøge vægten (i kg) af seks hvidkålhoveder, kunne se ud som:

$$X = (1.1, 1.3, 0.9, 1.5, 1.0, 1.1)$$

Her bemærkes at alle elementerne tilhører datatypen af talværdier med to decimaler.

Et sidste eksempel på et datasæt, der f.eks. kunne fremkomme ved at undersøge sammenhængen mellem højde og vægt for fem personer, kunne se ud som:

$$X = ((175, 85.2), (161, 68.7), (193, 102.4), (186, 92.8), (178, 79.4))$$

Her er datatypen altså talpar bestående af et positivt heltal samt et positivt decimaltal med en decimal.

Sorterede datasæt

Ud fra et givet datasæt kan man opstille et tilsvarende *sorteret datasæt* hvor der er flyttet rundt på dataværdierne så den mindste dataværdi står først og den største står sidst i datasættet.

Et eksempel på en sortering af et datasæt hvor datatypen er simple talværdier kan se ud som:

$$X = (2, 3, 3, 1, 4, 5, 4, 3, 2, 4) \quad \Rightarrow \quad X_s = (1, 2, 2, 3, 3, 3, 3, 4, 4, 4, 5)$$

Et eksempel på en sortering af et datasæt hvor datatypen er talpar kan se ud som:

$$X = ((3, 6), (1, 2), (3, 1)) \quad \Rightarrow \quad X_s = ((1, 2), (3, 1), (3, 6))$$

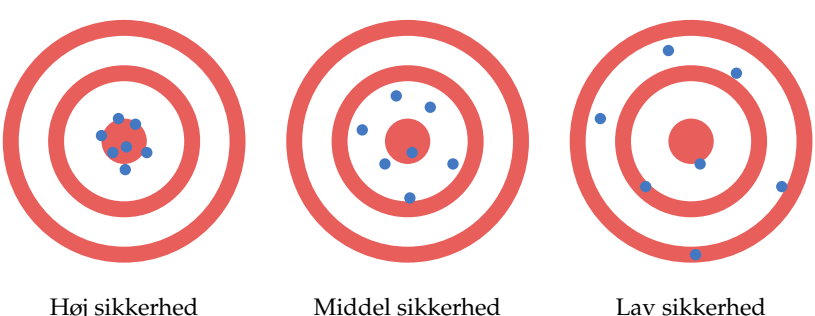
Her er sorteringen sket primært ift. den første talværdi i talparret og sekundært ift. den anden talværdi i talparret.

Fejlkilder ved målinger

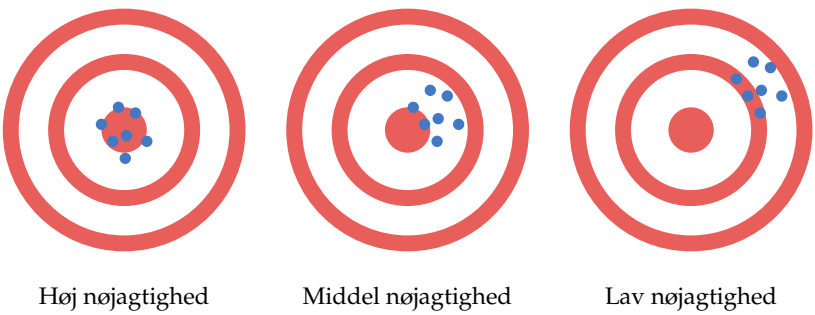
Data består som regel af enten optællinger eller målinger, der alt efter optællings- eller målemetoden er udsat for fejlkilder, der påvirker dataværdierne. Fejlkilderne kan opdeles i tre forskellige slags: målingens *sikkerhed*, *nøjagtighed* og *præcision*.

Når man taler om en målings sikkerhed, eller måske oftere omtalt som måleusikkerhed, menes der fejlkilder, som mere eller mindre skyldes tilfældigheder. Dette vil sige, at hvis man prøver at foretage den samme måling to gange vil måleusikkerheden gøre, at de to målinger ikke giver helt den samme talværdi. For at illustrere begrebet om målesikkerhed, betragtes figuren nedenfor hvor der er vist en skydeskive med punkter, der repræsenterer gentagne forsøg på at måle midten af skydeskiven. Når målesikkerheden er høj vil der være en lille spredning af punkterne og når målesikkerheden er lav vil der være en stor spredning.

Punkterne er spredt tilfældigt omkring midten af skydeskiven hvilket gør at man kan reducere en måleusikkerhed af en dataværdi ved at gentage en måling flere gange og derefter bestemme dataværdien som gennemsnittet af flere målinger.



Den næste type fejlkilde er *nøjagtighed*. I modsætning til målesikkerheden skyldes en målenøjagtighed en systematisk fejl, som påvirker alle målinger ens. På denne måde vil en lav målenøjagtighed resultere i, at målepunkternes centrum vil afvige fra skydeskivens midtpunkt som vist i figuren nedenfor.



Den sidste type fejlkilde er *præcision*. Målepræcision er mere eller mindre antallet af betydende cifre som dataværdien er angivet med. Jo mere man afrunder en måleværdi desto mere begrænser man en målings mulige værdier hvilket udgør en fejlkilde. Ift. analogien med skydeskiven kan målepræcision illustreret ved at målepunkterne kun kan ligge i krydspunkterne af et net eller koordinatsystem. Jo lavere præcision desto mere grovmasket net hvilket gør, at en måling potentiel bliver flyttet en større afstand fra dets virkelige værdi.

