STK-IN1050 - Statistikk for informatikere

1. Innledning & utforskende dataanalyse

Thordis L. Thorarinsdottir

2026-01-19

Struktur

- Forelesninger
  - Presentasjon av hovedinnhold, innledende eksempler og illustrasjoner
  - Mandag 14:15-16 og onsdag 12:15-14
  - Opptak av forelesningene legges ut i Canvas eller på semestersiden
- Åpne grupper
  - Hjelp til å løse ukesoppgaver, gjennomgang av noen oppgaver i plenum
  - Tirsdag 14:15-16 eller onsdag 10:15-12

Studenter som deltar aktivt i forelesninger og grupper bidrar til et dynamisk læringsmiljø og oppnår ofte større læringsutbytte!

Eksamen og obligatoriske oppgaver

- Eksamen
  - Midtveiseksamen 16. mars kl 15-17: teller 1/4 av endelig karakter
  - Slutteksamen 2. juni kl 9-13: teller 3/4 av endelig karakter
  - Hjelpemidler på eksamen: Godkjent kalkulator og formelsamling for STK-IN1050
- Obligatoriske oppgaver
  - Fire obligatoriske oppgaver med frist
    - Uke 5: Tirsdag 27. januar eller onsdag 28. januar
    - Uke 9: Torsdag 26. februar kl 14:30
    - Uke 16: Torsdag 16. april kl 14:30
    - Uke 18: Torsdag 30. april kl 14:30

Alle fire obligatoriske oppgaver må være godkjent før avsluttende eksamen.

Obligatorisk oppgave 1

- Møt opp på åpen gruppe tirsdag 27. januar kl 14:15 eller onsdag 28. januar kl 10:15
- Få registrert oppmøte hos gruppelærer
- Lever til gruppelærer et håndskrevet notat på maks 1/2 side hvor du skisserer hva du ønsker å få ut av dette kurset

Pensum

- Egen bok for emnet: https://pages.github.uio.no/thordist/SFoML/
- Forelesningsfoiler
- Ukesoppgaver

Alt pensum blir gjort tilgjengelig på semestersiden

Programvare

- Vi bruker Python for å illustrere statistiske metoder
- Innføring i Python programmering for praktisk data-analyse gis i forelesningen på onsdag 21. januar
- Vi bruker jupyterhub.uio.no til å illustrere eksempler og oppgaver

Kort om emnet

Etter å ha fullført emnet vil du:

- kjenne til de viktigste metodene for dataanalyse
- kunne velge den/de riktige metoden(e) for de tilgjengelige dataene
- forstå hvordan og hvorfor de valgte metodene (ikke) fungerer
- kunne komme med statistiske utsagn om populasjonens sammensetning og egenskaper, selv om bare et utvalg er tilgjengelig
- kunne analysere enkle statistiske sammenhenger
- kunne begrunne en valgt tilnærming til dataanalyse og tolke resultatene

Eksamenskrav

På eksamen skal du vise at du

- har forstått kursinnholdet
- kan velge en passende metode for å løse hver oppgave
- kan anvende metodene vi har diskutert
- kan tolke resultatene
- kan begrunne din tilnærming
- har mestret matematisk korrekt notasjon

Hva er statistikk?

Statistikk er vitenskapen om å samle inn, organisere, analysere og tolke numeriske fakta, som vi kaller data.

Hva er statistikk?

Statistikk er vitenskapen om å samle inn, organisere, analysere og tolke numeriske fakta, som vi kaller data.

"Statistics is about taking these different kinds of data that we might have, examining those data and opening those data up to tell the stories that the data reveal."

– Kerry Mengerson

https://www.youtube.com/watch?v=jbkSRLYSojo

“The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.”

– John Tukey

Dataanalyse i seks trinn



Dataanalyse i seks trinn



Vi vil i dette kurset fokusere på trinn 4 (som kan inkludere 3), 5 og den første delen av 6.

Data

Vi vil diskutere analyse av **kvantitative data**. Dette er data som foreligger i form av tall eller andre mengdetermer, i motsetning til **kvalitative data**, som vanligvis uttrykkes i form av tekst.

Kvantitative data kan ha forskjellige **målenivåer**:

- **Nominalnivå**: De mulige verdiene er forskjellige etiketter; har ingen rekkefølge og kan ikke meningsfullt tolkes som tall.
- **Ordinalnivå**: De mulige verdiene kan ordnes etter størrelse (større eller mindre), men intervallene mellom verdiene er ikke meningsfullt tolkbare.
- **Kardinalnivå**: De mulige verdiene kan tolkes som tall; intervallene (og vanligvis også forholdstallene) mellom verdiene er tolkbare.

Eksempler av data på forskjellige nivåer

- **Nominalnivå**:
  - Kjønn
  - Nasjonalitet
  - Formålet med et lån
- **Ordinalnivå**:
  - Skolekarakter
  - Rangering i et mesterskap
  - Kundetilfredshet
- **Kardinalnivå**:
  - Alder ved fullført utdanning
  - Høyde
  - Årsinntekt

Målenivået bestemmer hvilken type analyse er mulig

Nivå	Frekvenser	Rang	Differanser/ Forholdstall
nominal	ja	nei	nei
ordinal	ja	ja	nei
kardinal	ja	ja	ja

Målenivået bestemmer hvilken type analyse er mulig

Nivå	Frekvenser	Rang	Differanser/ Forholdstall
nominal	ja	nei	nei
ordinal	ja	ja	nei
kardinal	ja	ja	ja

Vi fokuserer hovedsaklig på data på ordinalnivå og kardinalnivå, men vi kan også være interessert i å modellere sammenhengen mellom f.eks. årsinntekt og kjønn.

Videre kategorisering av datatyper

Vi skiller også i mellom **numeriske** og **kategoriske** kvantitative data:

- **numeriske** data: informasjonen er registrert som tall, kan være
  - **kontinuerlig** (langs en sammenhengende skala; kardinalnivå)
  - **diskret** (angis som regel som heltall; ordinalnivå)
- **kategoriske** data: informasjonen faller i forhåndsdefinerte kategorier, kan være på
  - nominalnivå
  - ordinalnivå

Hvorfor lage grafiske fremstillinger av dataene?

- Visuell forståelse
- Identifisere avvik
- Informasjon om datafordeling/modell
- Identifisere relasjoner mellom variablene
- Identifisere datakvalitetsproblemer

Frekvenser og frekvensfordelinger

Anta at vi har observert numeriske data $x_1, \ldots, x_n$ som beskriver en bestemt **karakteristikk** X hos forskjellige **individer**, hvor n betegner **utvalgsstørrelsen**.

La $a_1, \ldots, a_J$ med $J \leq n$ være de mulige verdiene for karakteristikken X.

For numeriske karakteristikker er disse tallene også ordnet, dvs. $a_1 < \cdots < a_J$.

Frekvenser og frekvensfordelinger

Absolutte og relative frekvenser av funksjonsverdiene:

- $h(a_j) = h_j$ **absolutt frekvens** av verdien a_j, dvs. antall x_i fra $x_1, \ldots, x_n$ med $x_i = a_j$
- $f(a_j) = f_j = h_j/n$ **relativ frekvens** av a_j

Da har vi

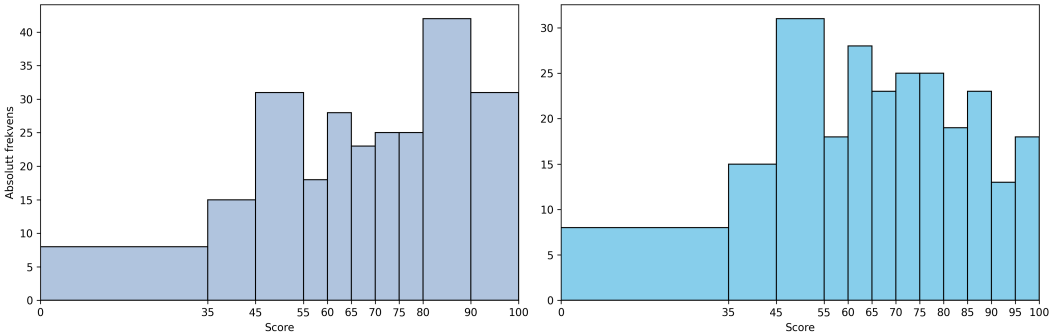
- Fordeling av de absolutte frekvensene: $h_1, \ldots, h_J$
- Fordeling av de relative frekvensene: $f_1, \ldots, f_J$

For kontinuerlige data med mange mulige verdier, bestemmes frekvensene vanligvis basert på en oppdeling av dataområdet i J-intervaller/klasser $[c_0, c_1], \ldots, (c_{j-1}, c_j], \ldots, (c_{J-1}, c_J]$.

Eksempel: Resultater fra midtveiseksamen i STK1000 H2025

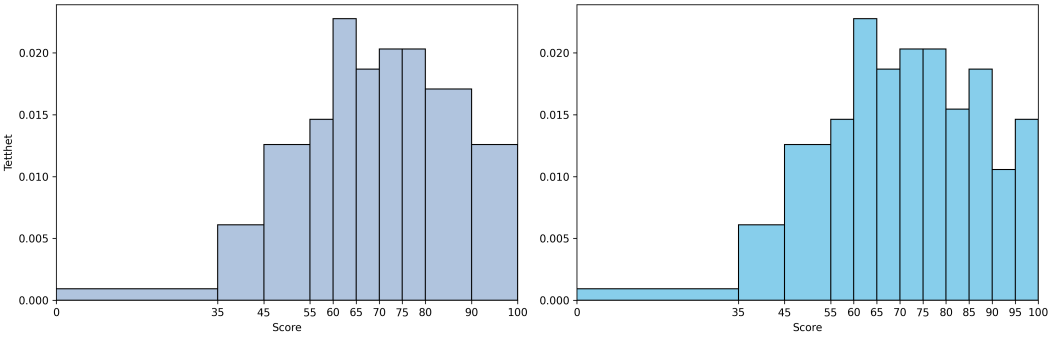
Score intervall	Absolutt frekvens	Relativ frekvens
[0, 35)	8	0.0323
[35, 45)	15	0.0605
[45, 55)	31	0.1250
[55, 60)	19	0.0766
[60, 65)	28	0.1129
[65, 70)	24	0.0968
[70, 75)	25	0.1008
[75, 80)	25	0.1008
[80, 90)	42	0.1694
[90, 100]	31	0.1250
Total	248	1.0000

Hvis vi plotter klassene mot absolutt frekvens i hver klasse, da avhenger den visuelle representasjonen av den valgte bredden på klassen:



Et annet alternativ er å la høyden på rektanglene være proporsjonal med

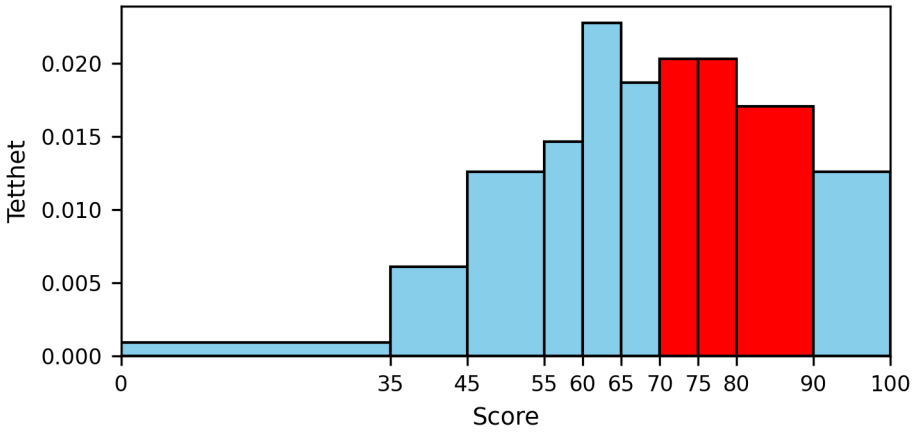
$$\text{tetthet} = \frac{\text{relativ frekvens}}{\text{bredde på rektanglet}}.$$



Tettheten introduseres formelt senere i kursen. Her er det nok å forstå at arealer under tettheten tilsvarer relative frekvenser.

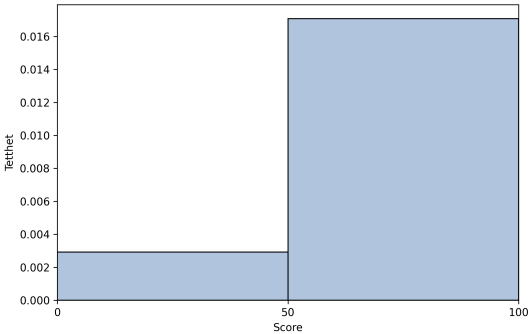
- Vi vil bruke histogrammer mye for visualisering av frekvenser av numeriske data, dvs. for visualisering av frekvensfordelingen.
- Det totale arealet av histogrammet er én, siden arealene tilsvarer andelene, og alle andelene til sammen er lik én.
- Følgelig kan deler av arealet brukes til å estimere frekvenser av enhver klasse.
- For eksempel, den estimerte andelen av studenter med score mellom a og b er gitt ved $\int_a^b H_x dx$, hvor H_x er høyden på histogrammet ved verdien x.

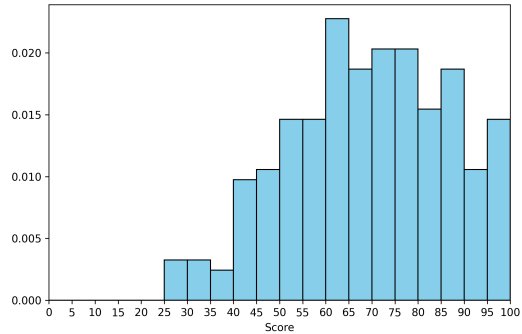
Eksempel: Andel studenter i STK1000 H2025 som fikk score mellom 70 og 90 på midtveiseksamen:



Vi bruker vanligvis L like store klasser for histogrammer.

Antallet L bør velges nøye $\Rightarrow$ **bias-varians-tradeoff**

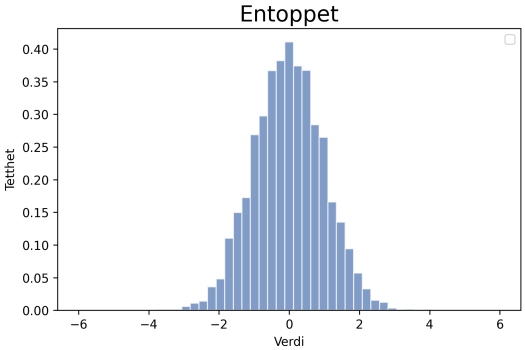


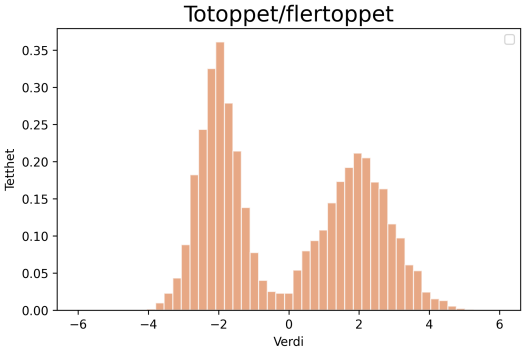


Klassifisering av frekvensfordelinger

Frekvensfordelinger kan klassifiseres i forskjellige typer basert på histogrammene.

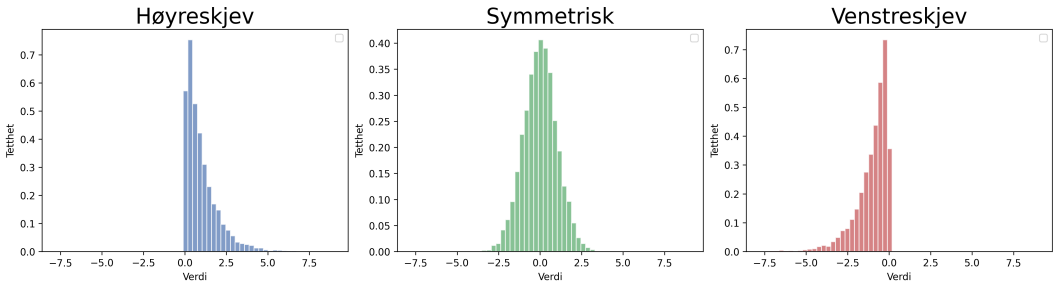
- Hvor mange topper har fordelingen?





Klassifisering av frekvensfordelinger

- Er fordelingen symmetrisk eller skjev?



Søylediagram (eller stolpediagram)

Hvis vi har kategoriske data, så kan vi bruke **søylediagram** til å vise fordelingen av verdiene i datamaterialet.

Her er de ulike kategoriene markert langs den horisontale aksen, og langs den andre aksen er det enten frekvenser (antall observasjoner av hver kategori) eller relative frekvenser (andelen observasjoner av hver kategori).

Sentralmål

Sentralmål eller et tall som beskriver hva som er vanlig eller typisk i et statistisk datamateriale:

1. **Gjennomsnittet** er gitt ved

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

2. **Medianen** deler observasjonene inn i 50% store og 50% små observasjoner:

$$x_{\text{med}} = \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{n oddetall} \\ \in \left[x_{\left(\frac{n}{2}\right)}, x_{\left(\frac{n}{2}+1\right)}\right], & \text{n partall.} \end{cases}$$

   hvor vi har ordnet observasjonene $x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n)}$.

Sentralmål

3. **Typetall** (eller **modalverdi**) x_{mod} er den observasjonen som forekommer flest ganger i et datamateriale.

Sentralmålene beskriver den sentrale tendensen til dataene:

- For gjennomsnittet gjelder at

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0,$$

  dvs. forskjellene til venstre og til høyre for gjennomsnittet kansellerer hverandre ut.

- For medianen ligger minst 50% av observasjonene til venstre og minst 50% av observasjonene til høyre for medianen.

Forutsetninger for bruk av sentralmål

- Gjennomsnitt:
  - gjelder for data på kardinalnivå
  - nyttig for data på ordinalnivå med to kategorier kodet som 0/1 (**binære** data)
  - noen ganger også brukt for andre data på ordinalnivå
- Median:
  - gjelder fra data på ordinalnivå og kardinalnivå
- Modalverdi:
  - gjelder for alle skalanivå
  - gir ikke alltid god mening for numeriske data med mange forskjellige mulige verdier
  - mindre relevant i det kurset

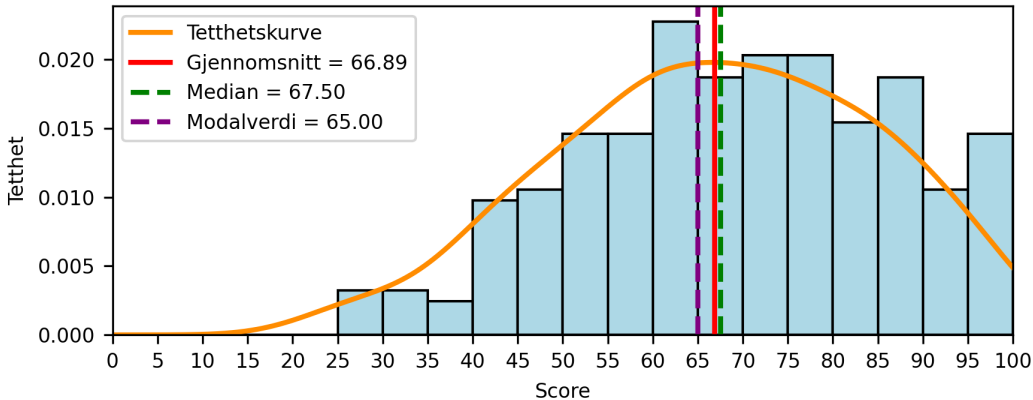
Sentralmål og form på frekvensfordelingen

For entoppede frekvensfordelinger gjelder:

- Symmetrisk fordeling:
$$\bar{x} \approx x_{\text{med}} \approx x_{\text{mod}}$$
- Høyreskjev fordeling:
$$\bar{x} > x_{\text{med}} > x_{\text{mod}}$$
- Venstreskjev fordeling:
$$\bar{x} < x_{\text{med}} < x_{\text{mod}}$$

For ekte data får vi ikke alltid et perfekt entoppet histogram; da kan disse reglene bryte sammen hva modalverdien angår.

Eksempel: Score på midtveiseksamen i STK1000 H2025



Kvantiler

Kvantiler er en generalisering av medianen: La $0 \leq \alpha \leq 1$. Et α-kvantil er et tall x_α der minst $\alpha \times 100\%$ av dataene er mindre enn eller lik x_α, og $(1 - \alpha) \times 100\%$ av dataene er større enn eller like x_α.

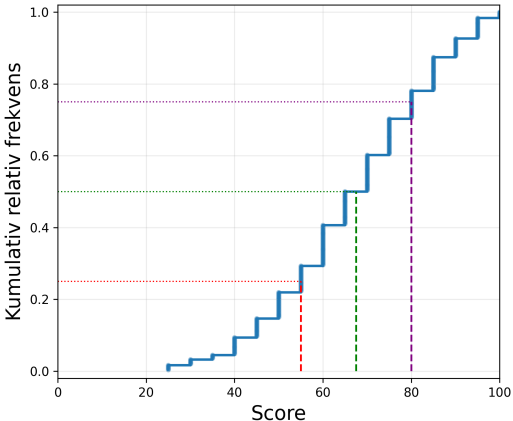
Vi kan finne kvantiler for data på ordinal- eller kardinalnivå.

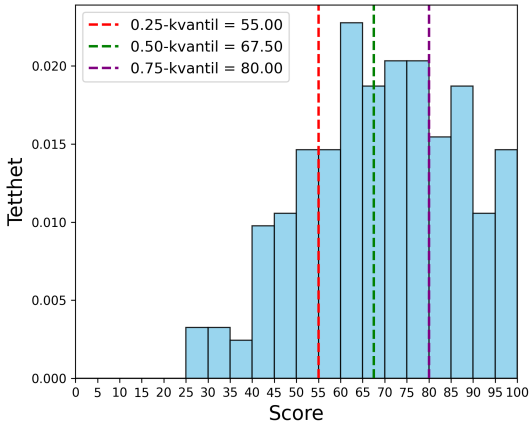
Vi bruker ofte

- Medianen: $x_{\text{med}} = x_{0.5}$
- Kvartilene: $x_{0.25}, x_{0.5}, x_{0.75}$
- Desilene: $x_{0.1}, x_{0.2}, \ldots, x_{0.9}$

Kvantilene kalles noen ganger også **persentiler** hvor medianen eller 0.5-kvantilen kalles 50-persentilen.

Eksempel: Score på midtveiseksamen i STK1000 H2025





Spredningsmål

Spredningsmål er et mål for spredningen av verdiene i et datamateriale.

- Den **empiriske variansen** s^2 og det **empiriske standardavviket** s måler spredningen i dataene rundt gjennomsnittet. De er gitt ved

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

og

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

(Vi diskuterer senere i kurset hvorfor vi dividerer med $n-1$ og ikke n).

Spredningsmål

- **Interkvartilavstand** (IQR) er et mål på spredning i et datamateriale, som viser området for de midterste 50% av dataene. IQR er gitt ved

$$\text{IQR} = x_{0.75} - x_{0.25}.$$

Eksempel: Score på midtveiseksamen i STK1000 H2025

```
import pandas as pd
pd.set_option("display.precision", 2)
df = pd.read_csv("../data/Midterm_STK1000.txt",sep='\s+', header=0)
df = df[['TSc']]

print(df.describe().T)
```

```
     count   mean    std   min   25%   50%   75%    max
TSc  246.0  66.89  17.45  25.0  55.0  67.5  80.0  100.0
```

Deskriptive statistikker: oppsummering

- sentralmål
  - gjennomsnitt
  - median
  - modalverdi
- spredningsmål
  - varians
  - standardavvik
  - interkvartilavstand (IQR)
- mål som avslører formen på en frekvensfordeling
  - kvantiler

Boksplott

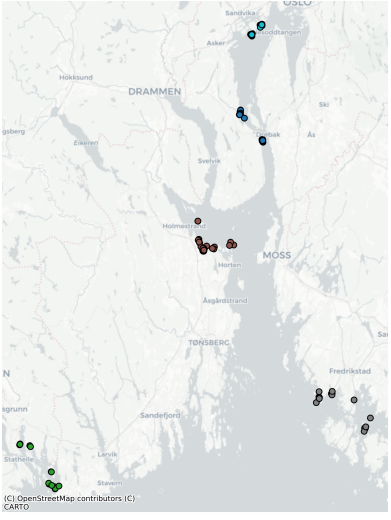
Boksplott er et grafisk **fem-punkts-sammendrag** av en frekvensfordeling som består av:

- minimumsverdien x_{min},
- nedre kvartil $x_{0.25}$,
- medianen x_{med},
- øvre kvartil $x_{0.75}$, og
- maksimumsverdien x_{max}.

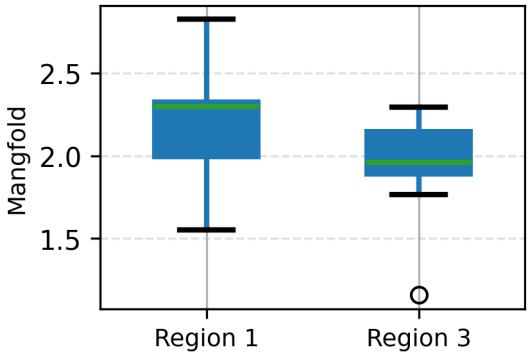
Eksempel: Biologisk mangfold i Oslofjorden

FRISK OSLOFJORD prosjektet har samlet inn data på biologisk mangfold i Oslofjorden. Prosjektet har utført tråling og basert på antall arter man har funnet, er det stedet blitt gitt en mangfoldsscore.

Kartet viser de lokasjonene hvor prosjektet har samlet inn data. Vi skal sammenligne dataene fra de to sørligste områdene.



Eksempel: Biologisk mangfold i Oslofjorden



Vi vil senere i kurset diskutere hvordan vi formelt kan konkludere om det er forskjell på biologisk mangfold i disse to regionene.

Boksplott: Oppsummering

Et boksplott gir en rask oversikt over formen på frekvensfordelingen (sentral tendens, variasjon, skjevhet, ekstremverdier).

Boksplotten viser **outliers** (dvs. usedvanlig små eller store observasjoner) separat.

Et datapunkt merkes som outlier i en boksplott hvis den har mer enn 1,5 IQR avstand fra $x_{0.25}$ eller $x_{0.75}$.

Oppsummering av dagens forelesning

- Data kan ha forskjellige målenivåer: målenivået bestemmer hvilken type analyse er mulig
- Vi starter en dataanalyse typisk med grafiske fremstillinger av datamaterialet for å skaffe oss en oversikt over de dataene vi skal jobbe med
- I tillegg bruker vi deskriptive statistikker til å oppsummere dataene

Neste forelesning

Neste forelesning onsdag 21. januar: Python programmering for data science med Geir Kjetil Sandve.