Morven Muilwijk
E-mail: morven.muilwijk@student.uib.no

UNIVERSITY OF BERGEN
GEOPHYSICAL INSTITUTE

Grupperte data

Vedlegg til Ekte data oppgave

January 13, 2016

Grupperte data
For å oppnå en mer oversiktlig og informativ fremstilling av et innhentet datasett, er det vanlig å gruppere tallmaterialet. Det skilles mellom to forskjellige typer variabler:

- **Diskrete** variabler: Observasjonene kan kun ha visse (diskrete) verdier som er adskilt fra hverandre (eks:antall øyne i et terningskast).
- **Kontinuerlige** variabler: Observasjonene kan ha hvilke som helst verdier innenfor et begrenset eller ubegrenset definisjonsområde (eks: tiden, t).

I prinsippet er det ofte en "glidende" overgang fra kontinuerlige til diskrete variabler. La oss som eksempel betrakte "pers'en" på 60 m til studenter ved Høgskulen i Bergen. I prinsippet er dette en kontinerlig variabel som kan ha hvilken som helst verdi mellom, la oss si, 6 s og 15 s. I praksis måles imidlertid tiden på nærmeste tidel eller hundredel. Vi har da med diskrete verdier å gjøre. Ved tilstrekkelig fin inndeling vil det ikke gi noe praktisk forskjell om vi betrakter data som kontinuerlige eller diskrete. Det samme gjelder måledata data.
Normal fremgangsmåte for gruppering av oseanografiske data bestående av enkelttall kan være:

1. **Bestemmelse av maksimum og minimum.**
   Vi finner minste og største observasjonsverdi, x_{min} og x_{maks}.
2. **Klasseinndeling**
   Vi deler x-området inn i k-*klasser*, som regel med like brede *klasseintervaller*. Klassene må ikke overlappe hverandre, og til sammen må klassene dekke alle verdier fra minimum til maksimum.
3. **(Frekvens-) tabell**
   Vi lager en tabell med god plass og mange kolonner. Første kolonne anngir de ulike klasseintervallene, med de lavest verdiene øverst og de største nederst. Hva som skal stå i de andre kolonnene kan være avhengig av oppgaven, men for eksempel:
   - *Klassemidtpunkt*, dvs midtpunkt mellom de egentlige klassegrensene.

Morven Muilwijk
E-mail: morven.muilwijk@student.uib.no

UNIVERSITY OF BERGEN
GEOPHYSICAL INSTITUTE

- *Tellekolonne*: Du merker av en strek i riktig klasserubrikk for hver av dine observasjoner.
- *Frekvens-kolonne*: Du oppgir antall observasjoner innenfor hver klasse. Dette kalles klassefrekvens.
- *Relativ frekvens-kolonne*: Klassefrekvensen delt på antall observasjoner.
- *Kumulativ frekvens-kolonne*: Sum av alle klassefrekvensene fra og med den første raden til og med den klassen du ser på. i siste klasse vil kumulativ frekvens alltid være lik antall observasjoner.
- *Relativ kumulativ frekvens-kolonne*: Du tar kumulativ frekvens og deler på antal observasjoner.
- $m_i * f_i$*-kolonne*: For hver klasse beregner du produktet av klassemidtpunkt og klassefrekvens. Hensikt: regne gruppert middelverdi.
- $m_i^2 * f_i$*-kolonne*: For hver klasse beregner du produktet av av kavadratet av klassemidtpunktet og klassefrekvensen. Hensikt: beregne gruppert standardavvik.

Følgende formler brukes for å finne gruppert middelverdi og gruppert standardavvik:

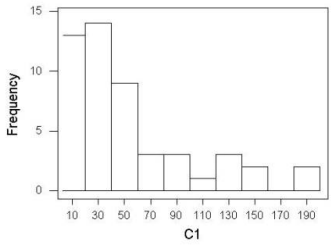
$$\text{Gruppert middelverdi}: \bar{x}_g = \frac{1}{n}\sum_{i=1}^{k} m_i * f_i \tag{0.1}$$

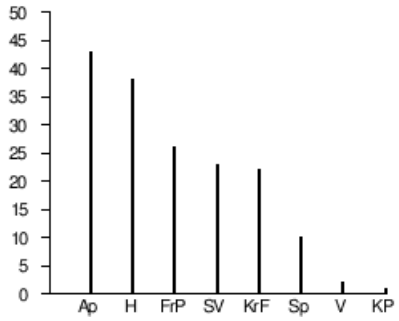
$$\text{Gruppert standardavvik}: s_g = \sqrt{\frac{1}{n}\sum_{i=1}^{k}(m_i - \bar{x}_g)^2 f_i} \tag{0.2}$$

Histogram og stolpediagram

Grupperte data kan enkelt fremstilles grafisk. For eksempel mye bruke er relativ frekvenshistogram eller relativ frekvens stolpediagram. Histogrammet består av rektangler: Et rektangel pr. klasse og rektangelareal lik relativ klassefrekvens. Bredden av hvert rektangel er lik klassebredden.

Relativ frekvens-stolpediagram benyttes kun for diskrete variabler. Da tegnes en loddrett stolpe for hver diskrete verdi der vi har observasjoner. Høyden på hver stolpe (y-verdien) er normalt lik relativ frekvens.





Morven Muilwijk
E-mail: morven.muilwijk@student.uib.no

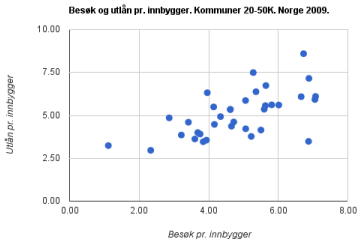
UNIVERSITY OF BERGEN
GEOPHYSICAL INSTITUTE

Empirisk korrelasjonskoeffisient og spredningsdiagram

Hittil har vi sett på enkelttal. Vi skal nå se på tallpar som for eksempel temperatur og saltholdighet sammen per måling. La oss belyse denne forskjellen ut fra tabellen nedenfor over samhørende 3MN og 3FY karakterer til 5 tilfeldig utvalgte studenter. Dataene er tegnet inn i et x-y diagram, som er et eksempel på et spredningsdiagram. Vi ser på karakterene som tallpar. Med tallpar kan vi belyse en del problemstillinger knyttet til hvordan x- og y-verdiene samvarierer. For eksempel:

- er det noen form for systematisk sammenheng mellom x- og y-verdiene?
- kan vi tallfest i hvor stor grad det er en slik sammenheng?
- kan vi tilpasse en fornuftig funksjon som beskriver sammenhengen mellom disse verdiene?
- kan vi forutsi en variabel dersom den andre er kjent?

Å studere disse variablene hver for seg vil ikke være til hjelp når det gjelder å svare på disse spørsmålene. Vi ønsker å finne korrelasjonen mellom de.

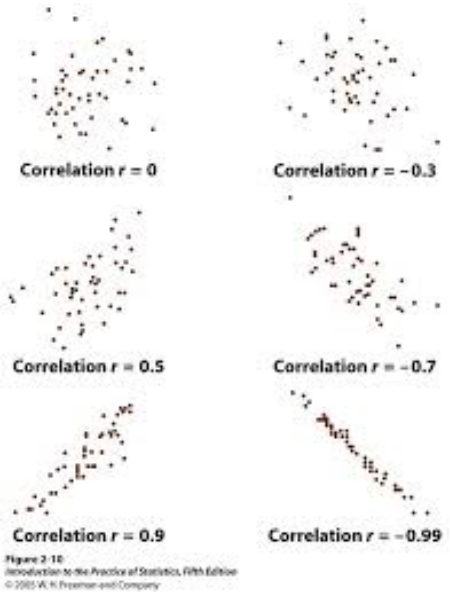


Den empiriske korelasjonskoeffisienten, som vi skal betegne med r, er et mål på graden av linær sammenheng melom x- og y-variablene. Før vi introduserer formelen skal vi angi noen viktige egenskaper og diskutere på hvilken måte den kan brukes til å måle graden av lineær sammenheng.

- r-verdiene ligger alltid mellom 1 og -1.
- Absoluttverdien indikerer graden av lineær sammenheng men fortegnet indikerer retning.
- $r > 0$ hvis mønsteret er et bånd som løper fra nedre venstre til øverste høyre hjørne.
- $r < 0$ hvis mønsteret er et bånd som løper fra fra øvre venstre til nedre høyre hjørne
- $r = 1$ hvis alle verdiene ligger eksakt på en og samme rette linje med et positivt stigningstall.
- en r-verdi nær null betyr liten grad av sammenheng. Se figur.

Morven Muilwijk
E-mail: morven.muilwijk@student.uib.no

UNIVERSITY OF BERGEN
GEOPHYSICAL INSTITUTE



- NB! det kan være lett å mistolke en observert korelasjon mellom to variabler som et årsaksforhold mellom variablene. Et klassisk eksempel er den positive korrelasjonen mellom antall storker og barnefødsler i store byer. Årsaken til dette er ikke at babyene kommer med storken, men at det er tredje variabel som spiller inn – nemlig størrelsen på byene. Jo større byer, jo flere storker, og jo større byer jo flere barnefødsler. Det er altså størrelsen som får antall storker og fødsler til å variere i samme retning.

Korrelasjonen kan regnes ut ved hjelp av følgende formel:

$$r = \frac{S_{xy}}{\sqrt{S_x^2}\sqrt{S_y^2}} \tag{0.3}$$

der

$$S_x^2 = \sum (x_i - \bar{x})^2 = \sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2 \tag{0.4}$$

$$S_y^2 = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2 \tag{0.5}$$

$$S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - \frac{1}{n}\left(\sum x_i\right)\left(\sum y_i\right) \tag{0.6}$$

og alle summer er fra 1 til n.