Valg av studieobjekt
– enheter og operasjonaliseringer

Silje Synnøve Lyder Hermansen

s.s.l.hermansen@stv.uio.no

Univariate analyser

Jacobsen (2005), kap. 14

Forberedelser i forkant – datastruktur

Datastrukturen i kvantitative analyser følger en spesifikk mal. Denne malen likner på et ruteark/en Excelfil. Når et slikt "ruteark" inneholder kolonnenavn og linjenummer, kaller vi dokumentet en "datamatrise" (i R heter denne kategorien `data.frame`).

- **Enheter:** Hver linje er viet til en enhet. Vi har like mange linjer som vi har enheter i datasettet. Hver enhet bruker også ofte å ha en id-kode som gjør det lettere å identifisere hver enhet. I spørreundersøkelser er dette nødvendig på grunn av personvern. I andre undersøkelser vil en standardisert ID gjøre det lettere å slå sammen datasett, og unngå problemer grunnet stavefeil (av personnavn, for eksempel).

Forberedelser i forkant – datastruktur (forts.)

- **Variabler:** Hver kolonne tilsvarer en variabel. Kolonnenavnet er det samme som variabelnavnet, og antall kolonner er det samme som antall variabler. Lengden på hver variabel er den samme som antall enheter. Det vil si at jeg i prinsippet må ha observasjoner av alle enhetene mine på den variabelen. Jeg kan ikke sammenlikne enheter som ikke er sammenliknbare.
- **Verdier:** Hver "rute" i rutearket tilsvarer en verdi. Hver variabel har dermed like mange verdier som vi har enheter. Når jeg mangler verdier, må jeg kode dem som manglende/"missing". I R gjøres dette ved å gi verdien `NA` ("not available"). Dette er hva Jacobsen kaller frafall av type 4.

Forberedelser i forkant – eksempel på datastruktur (forts.)

I en spørreundersøkelse fra 2006 ble respondentene spurt hvilket parti de stemte ved sist valg, samt en rekke andre spørsmål. Datasettet inneholder 380 enheter (linjer) og 10 variabler (kolonner). Det følgende er en smakebit fra datasettet:

	PARTIVLG	skyldigfri	merpoliti	burinne
3	Fremskrittspartiet (FrP)	0	1	2
5	Fremskrittspartiet (FrP)	0	2	2
6	Fremskrittspartiet (FrP)	0	1	1
8	Fremskrittspartiet (FrP)	0	2	3
10	Fremskrittspartiet (FrP)	0	0	1
13	Fremskrittspartiet (FrP)	0	1	2

Forberedelser i forkant – koding av variabler

Kvantitative dataanalyser baserer seg på mange enheter og få variabler. For at enhetene skal være sammenliknbare, må vi standardisere verdiene. Det innebærer å gi en tallverdi til metriske (og eventuelt ordinale) variabler, og å lage ferdige kategorier for kategoriske variabler (rette skrivefeil for eksempel). Av og til slår vi også sammen kategorier.

- Jo høyere målenivået er, jo flere analysemetoder har vi til rådighet. Samtidig trenger vi variabler som er realistiske.
- Alle variabler kan behandles som kategoriske, men kategoriske variabler kan ikke behandles metrisk.

Univariate analyser

Kvantitative analyser starter alltid med univariate analyser: Vi utforsker én variabel av gangen. Det er tre ting vi alltid vil være interessert i:

- **Fordelingen på variabelen:** Hvor mange enheter har hvilke verdier?
- **Den typiske verdien(e)** for enhetene.
- **Spredningen:** Det reelle verdispennet i variabelen.

Måten vi analyserer disse tingene på avhenger av målenivået på variabelen.

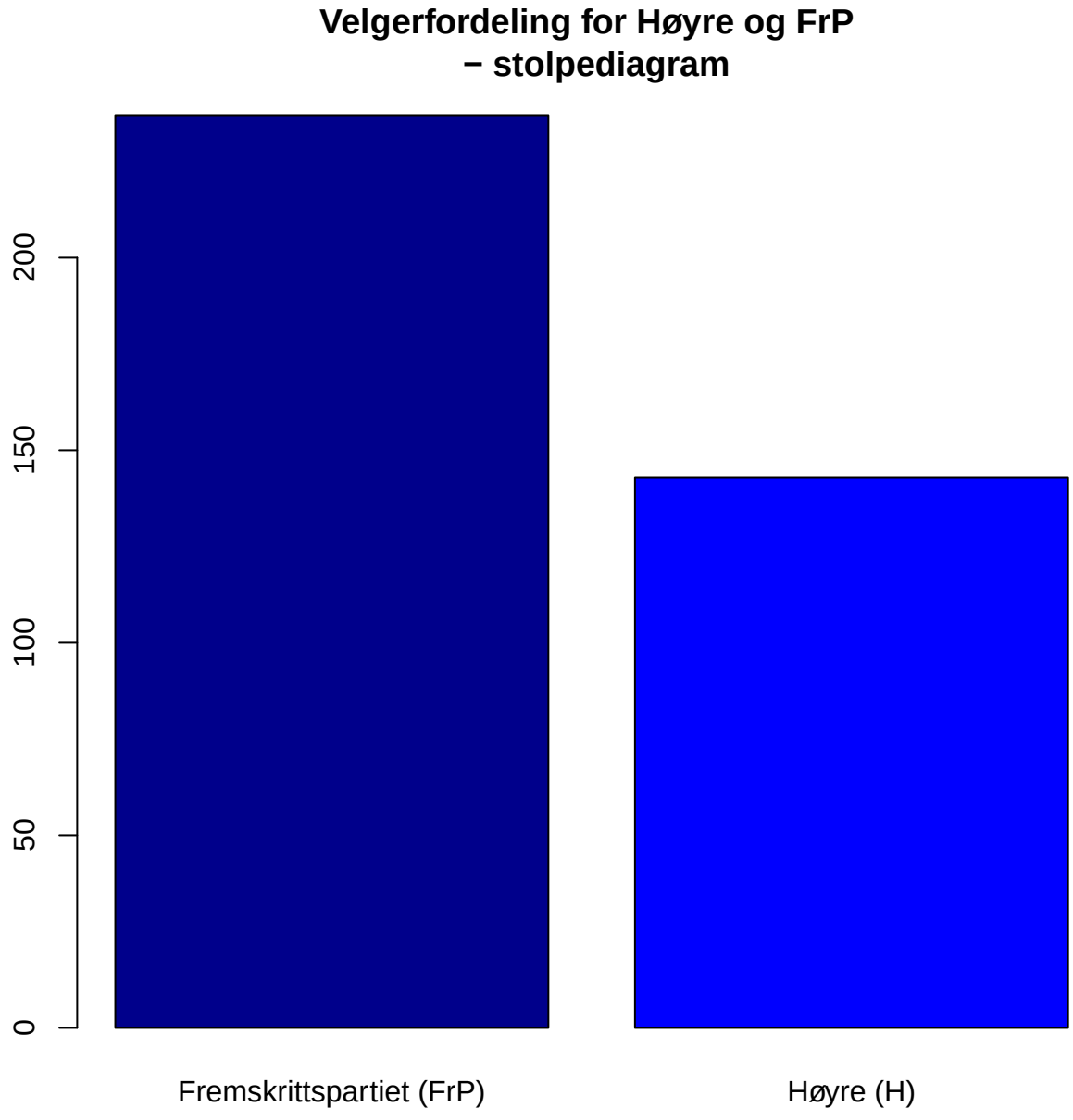
Kategoriske variabler

Jacobsen (2005), kap. 14

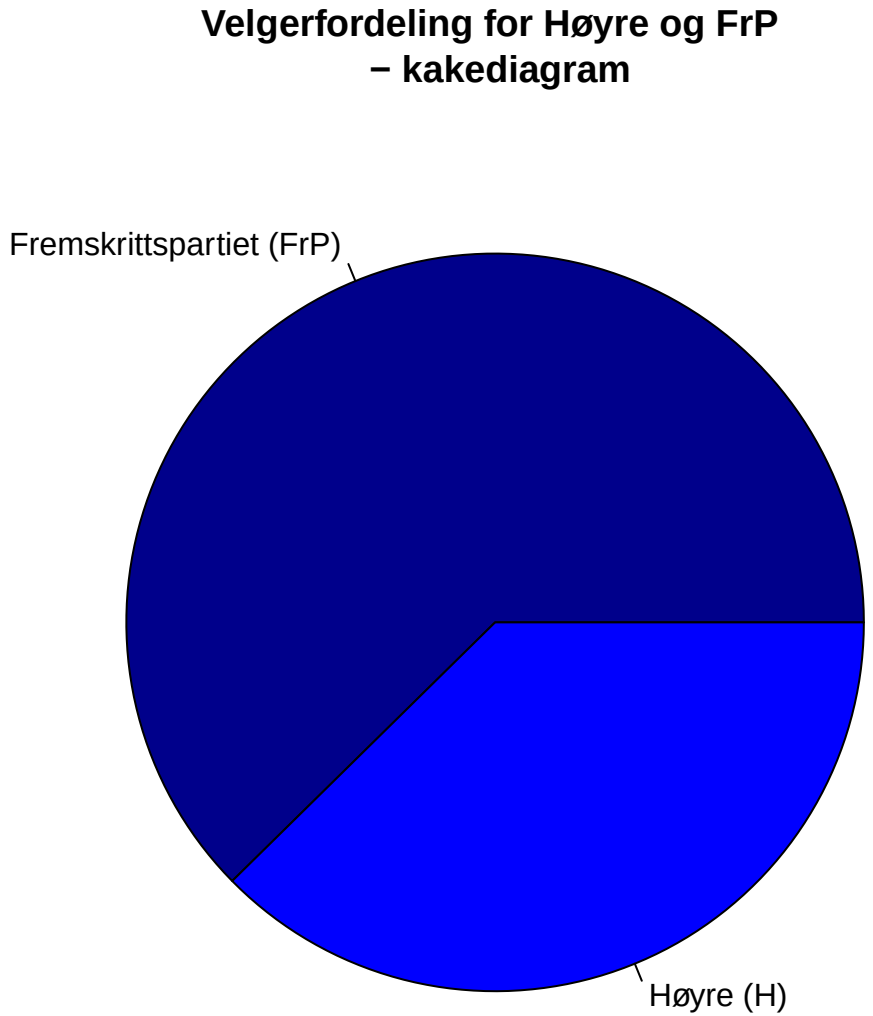
Kategoriske variabler

Kategoriske verdier (nominale verdier): Vi kan bruke disse til å uttale oss om likheter og forskjeller. I R kalles det nominale målenivået `factor` og kategoriene kalles `levels`. Disse kan kodes som tall (det interne løpenummeret i R), men tallene er ikke reelle tall (2 er ikke større enn 1). Eventuelt kan vi bruke kategorinavn. Da må kategoriene skrives helt likt, uten skrivefeil. Eks.: `"Fremskrittspartiet (FrP)"` og `"Fremskrittspartiet (FrP) "` vil i R ansees som to ulike kategorier.

Kategoriske variabler – fordelinger (visuell analyse)



Kategoriske variabler – fordelinger (visuell analyse)



Kategoriske variabler – Fordelinger (tabell)

Vi kan vise fordelingen i absolutte tall eller relative tall:

- **Absolutte tall:** Ligger alltid til grunn. Vi starter alltid med dette.
- **Relative tall:** Gjør fremstillingen ryddigere (særlig når vi har mange enheter), og åpner for sammenlikninger. De to vanligste enhetene er prosenter og proporsjoner:
  - $Proporsjon = \frac{Antallikategori}{Total}$
  - *Prosent* For å oppnå prosent, multipliserer vi med hundre.

	Absolutte tall	Proporsjon	Prosent
Fremskrittspartiet (FrP)	237.00	0.62	62.37
Høyre (H)	143.00	0.38	37.63

Table : Velgerfordeling for Høyre og FrP – tabell

Kategoriske variabler – typiske verdier

Som forskere er vi opptatt av hva som er *typisk* for ulike enheter. Det er slik vi også finner ut hva som ikke er typisk/ hva som skiller seg ut. Når målenivået er kategorisk(nominalt), kan vi kun bruke **modus** for å uttrykke hva den mest typike trenden er blant enhetene.

- **Modusverdi** er hvilken kategori som har flest verdier.
- **Modalprosenten** er prosentandelen av enhetene som befinner seg i denne kategorien.

Her er modusverdien "Fremskrittspartiet", og modalprosenten er på 62.

Kategoriske variabler – spredning

- Jo høyere **modalprosenten** er, jo mindre spredning er det i fordelingen.
- Jo flere kategorier man har, jo mindre kan den maksimale modalprosenten være. Derfor opererer man av og til med en **normert modalprosent**som varierer mellom 0 og 1.

Rangordnede variabler (ordinalt målenivå)

Jacobsen (2005), kap. 14

Ordinale variabler

Ordinale variabler (rangordnede kategorier): betyr at vi både kan skille kategorier av enheter fra hverandre, og at kategoriene kan rangeres i stigende rekkefølge. Kodeproblemer oppstår stort sett når vi befinner oss i grenselandet mellom ordinale- og kategoriske eller tellevariabler.
I R er de ordinale variablene kategorisert som `factor`, men vi kan velge å rangere/rydde faktornivåene (`levels`) for hånd etter et substansielt kriterium i stedet for alfabetisk slik R gjør.

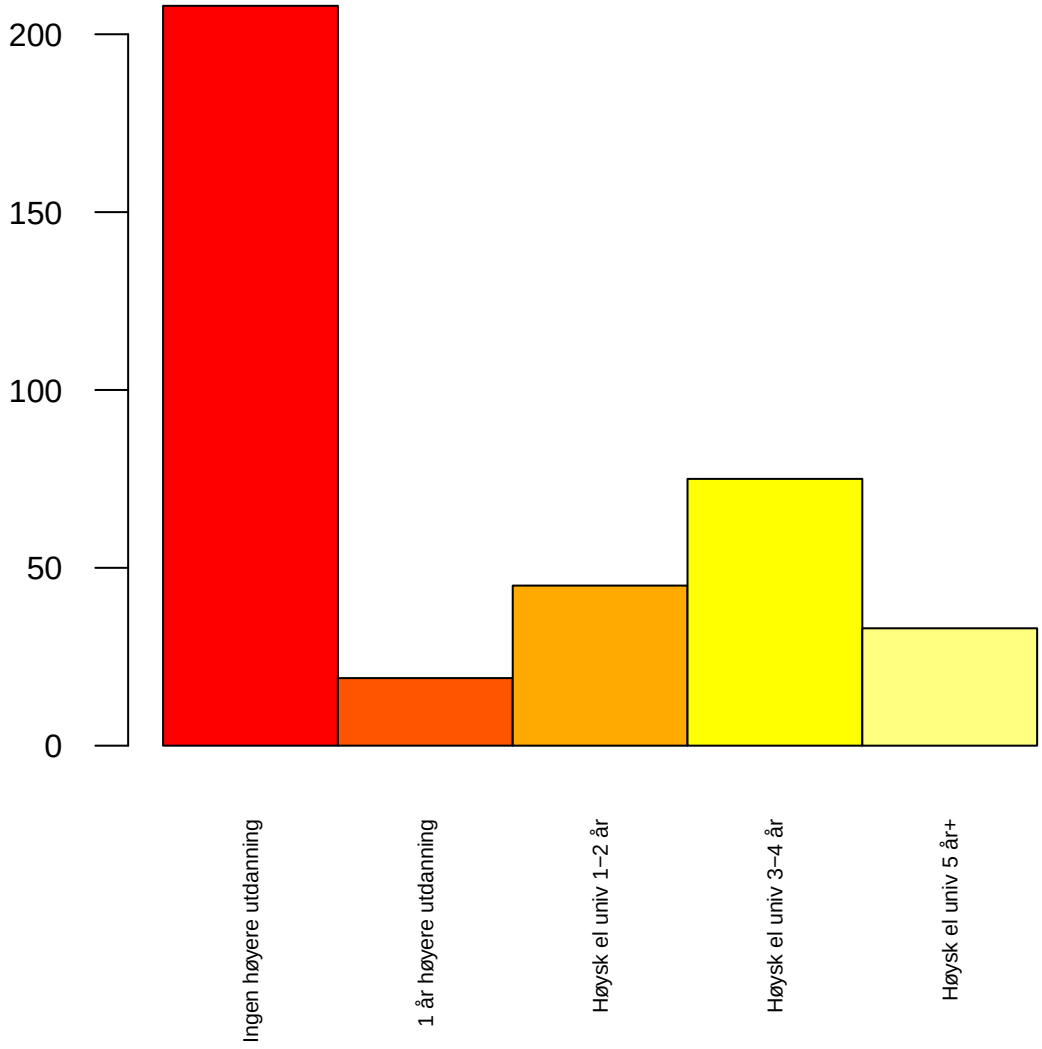
Ordinale variabler – fordeling

Et eksempel fra datasettet er variablen `utdanning`.
Respondentene er plassert i fem kategorier med løpenummer 1 til 5. Vi kan diskutere om 5 kategorier eventuelt er en tellevariabel (vi kan for eksempel regne snittet) eller en kategorisk variabel.

	Kategorinavn	Frekvens	Proporsjon
1	Ingen høyere utdanning	208	0.55
2	1 år høyere utdanning	19	0.05
3	Høysk el univ 1-2 år	45	0.12
4	Høysk el univ 3-4 år	75	0.2
5	Høysk el univ 5 år+	33	0.09

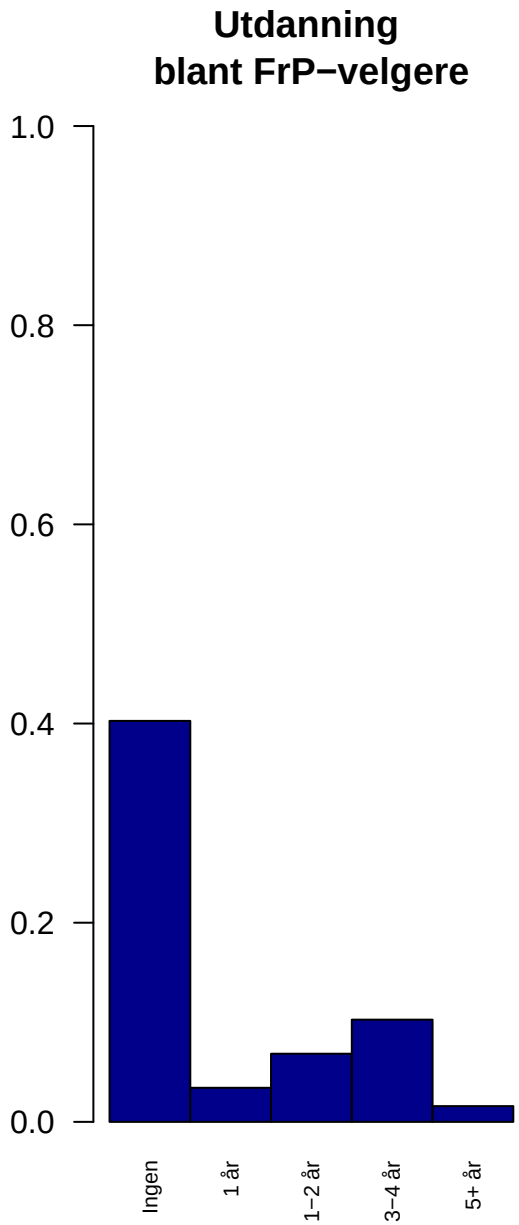
Velg å rangere ordinale variabler på en intuitiv skala: Høye verdier bør her være høy utdannelse.

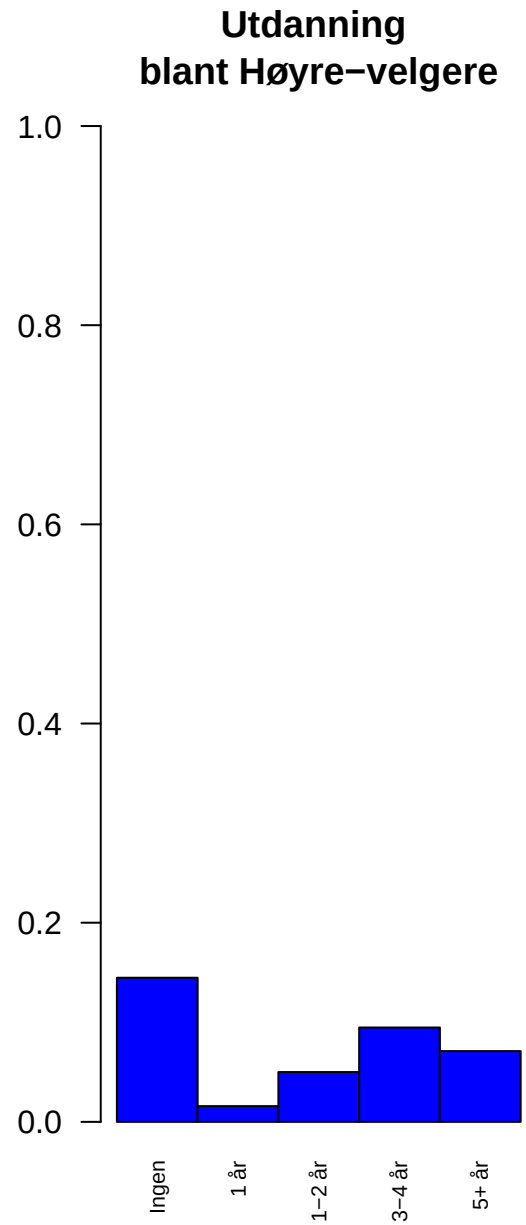
Frekvensfordeling over utdanning blant Høyre og FrP−velgere



Ordinale variabler – fordeling

Med flere kategorier kan kakediagram være rotete. Når variabler er på målenivå med ordinale verdier eller mer, er det også interessant å se på frekvensfordelingen. Samler fordelingen seg rundt midten? I dette eksempelet ser vi at de færreste respondentene har høyere utdanning.


Utdanning
blant FrP−velgere
1.0
0.8
0.6
0.4
0.2
0.0
Ingen
1 år
1−2 år
3−4 år
5+ år


Utdanning
blant Høyre−velgere
1.0
0.8
0.6
0.4
0.2
0.0
Ingen
1 år
1−2 år
3−4 år
5+ år

Ordinale variabler – fordeling

Når vi bruker den relative (proporsjonale) fordelingen, er det lettere å sammenlikne fordelingen internt i undergrupper: Dette er begynnelsen på en bivariat analyse.
I dette eksempelet ser vi at FrP-velgerne gjevnt over har lavere utdanning enn Høyre-velgerne.

Ordinale variabler – typiske verdier

Når målenivået er ordinalt, kan vi i tillegg til å bruke modus, forholde oss til **medianen** for å uttrykke hva den mest typike trenden er blant enhetene. Medianen uttrykker midtpunktet i en rangert fordeling: det er punktet som deler enhetene i to like store grupper: Når enhetene er ordnet etter verdi, vil enhet nummer $\frac{antall}{2}$ eventuelt $\frac{antall+1}{2}$.

1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 [38] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 [75] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 [112] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 [149] 1 1 1 1 1 1 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 [186] 1 1 1 1 1 1
1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 [223]
2 2 2 2 2 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
3 3 3 [260] 3 3 3 3 3 3 3 3 3 3 3 3 3 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
4 4 4 4 4 4 4 4 4 [297] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 4
4 4 4 4 4 4 4 4 4 4 4 4 4 4 4 [334] 4 4 4 4 4 4 4 4 4 4 4 4 4 4 5 5
5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 [371] 5 5 5 5 5 5 5 5 5 5

Ordinale variabler – typiske verdier

I dette tilfellet er enhet nummer $\frac{380}{2} = 190$ den som definerer medianen. Jeg kan lese av på linjenummer 190 hvilket utdannelsesnivå som utgjør medianen. I vårt eksempel er medianverdien på utdannelsesvariabelen 1 år.

Ordinale variabler – spredning

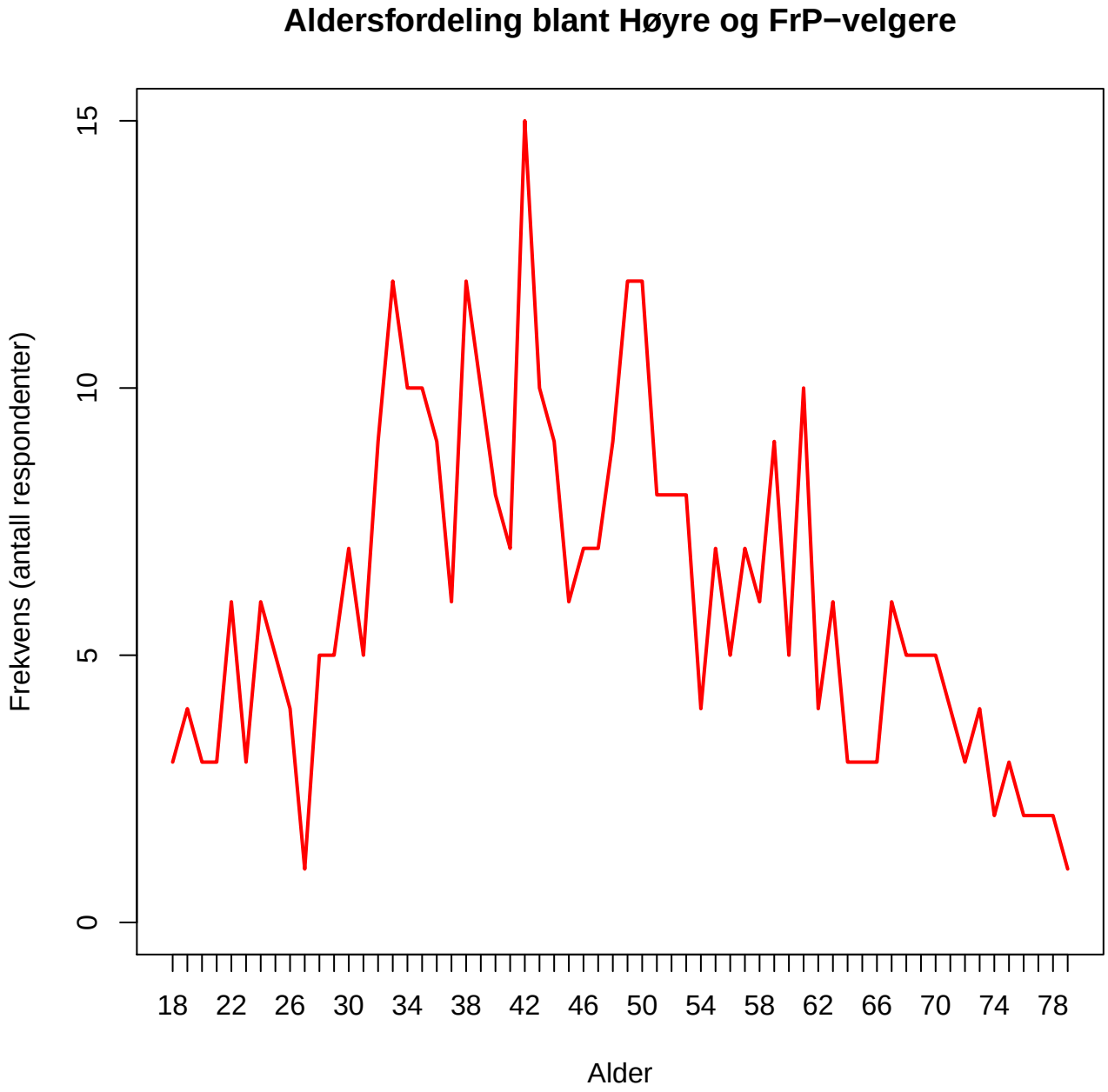
I følge Jacobsen kan vi bruke alle typer mål for spredning på ordinale variabler. Det forutsetter at vi omkoder og behandler dem som tellevariabler. Men skal vi behandle dem som ordinale variabler, har vi – i tillegg til modalprosenten – ytterligere et mål for spredning: **Minimumsverdien** og **maksimumsverdien** i fordelingen.
I vårt utdanningseksempel vet vi at minimum utdanning er "Ingen høyere utdanning" og maksimum er "5 år eller mer".

Tellevariabler

Jacobsen (2005), kap. 14

Tellevariabler

Tellevariabler (numeriske variabler): I R heter disse `numeric`, og er de eneste variablene vi kan gjøre matematiske utregninger på. Når vi ønsker å endre målenivå i R, bør vi gå gjennom et siste målenivå (`character`), bokstavnivå, før vi endrer noe til å bli numerisk. Det er fordi R alltid forholder seg til de interne løpenumrene (eks: 1 til 5), og ikke kategorinavnene (”1 år”, ”1-2 år” osv.).


Aldersfordeling blant Høyre og FrP−velgere
Frekvens (antall respondenter)
0
5
10
15
18
22
26
30
34
38
42
46
50
54
58
62
66
70
74
78
Alder

Tellevariabler – typiske verdier

Med tellevariabler kan vi også regne ut **gjennomsnittet** som et mål på en sentraltendens/typisk verdi. Når frekvensfordelingen er høyest på midten og lavest i ytterkantene, vil snittet likne på medianen og modus. Når fordelingen er sjev, slik vi så for utdannelsesvariabelen, vil disse være ulike.

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
Alder	18.00	35.00	45.00	46.17	57.00	79.00

Tellevariabler – spredning

For tellevariabler har jeg ytterligere to spredningsmål:

- **Variasjonsbredde:** differansen mellom maksimums- og minimumsverdien i fordelingen. I vårt eksempel er variasjonsbredden i alder avstanden mellom eldste og yngste respondent: $79 - 18 = 61$ år.
- **Standardavvik ("standard deviation", *sd*):** er et mer teknisk mål. Målet er definert som den kvadrerte differansen mellom gjennomsnittet i variabelen og hver enkelt enhet. Vi tar så gjennomsnittet av disse gjennomsnittene før vi tar roten av svaret. Vi kan få R til å gjøre dette for oss. I vårt eksempel er standardavviket på 15 år.

Spredningen sier noe om hvor "flat" frekvensfordelingen vår ser ut; hvor stor forskjell det er på enhetene internt i variabelen. I en høy smal fordeling/en fordeling med lav spredning er de fleste enhetene omtrent like.

Beslutningstre for ulike utvalgsmetoder