

P2 - Simulering af Data

Gruppe B2-19

2020-02-27

Contents

Chapter 1

Introduktion

Chapter 2

Variabeltyper

2.1 Kvalitative variabler

De stokastiske variabler kan opdeles i kategoriske eller kvalitative variabler. Ved en kvalitativ variabel fremgår der en enhed, som er den enhed der kategoriseres efter. Udfaldsmængden er de mulige værdier, som kan forekomme ved en undersøgelse af enheden. Dette kunne eksempelvis være nationalitet, hvor der kategoriseres efter land, hvor udfaldsmængden er landene i verden.

De kvalitative variabler er opdelt efter en nominal og ordinal skala.

Den nominale beskriver variabler, som ikke har en naturlig rangorden, og derved ikke kan rangeres. Ofte er de nominale variabler tekststrengene, men de kan også være numeriske i form af postnumre, dog vil det være meningsløst at foretage beregninger såsom middelværdi eller standardafvigelsen. Eksempler på nominale variabler er køn, hudfarve, nationalitet og hårfarve, da udfaldsmængden i enheden ikke kan rangordnes. Derved skal nominale variabler illustreres ved et cirkeldiagram, da der ikke er givet nogen skala. De udregninger, som der kan foretages på nominale variabler er andele og frekvenser, da disse kun kan skelne udfald.

De ordinale variabler er de kvalitative variabler, som kan rangeres. Eftersom udfaldsmængden er naturligt rangeret, og samtidig er kategorisk, vil det altså være en ordinal variabel. Eksempler på ordinale variabler er økonomisk status, uddannelsesgrad og karakterer. Disse variablers udfaldsmængde er altså rangeret i forhold til hinanden, da det er bedre at være i en høj indkomst fremfor en lav indkomst og ligeledes er det bedre at få karakteren 10 end karakteren 4. Ud fra observationerne kan man efterfølgende blandt andet finde medianen, kvartiler m.m.

Ønsker man at finde middelværdien skal man være lidt påpasselig, eftersom det reelt set ikke giver nogen mening på ordinale variabler. Dog er det muligt at

udregne en middelværdi, hvis der er en fælles forståelse af, hvad middelværdien betyder for enheden og hvilken skala, det er udregnet på baggrund af. Når ordinale variabler illustreres, skal det gøres som et søjlediagram efter rangordenen.

(?, s. 12-13)

2.2 Kvantitative variabler

Kvantitative variabler er defineret ved en numerisk værdi, hvilket gør variablerne målbar. Da en kvantitativ variabel antager en numerisk værdi, kan aritmetriske operationer udføres på variablen. På baggrund af disse simple operationer er det muligt at få informationer ud af et datasæt, der kan skabe indsigt i forhold til fokusområder, som skal undersøges yderligere. Dette er en væsentlig del af dataanalyse, således at ressourcer kan anvendes optimalt.

Et andet essentielt værktøj i dataanalyse er visualisering. På baggrund af middelværdi, varians og standardafvigelse, er det muligt at visualisere data med eksempelvis et histogram eller et boksplot.

Kvantitative variabler kan opdeles i to underkategorier, interval og ratio. Derudover kan disse underkategorier inddeles yderligere i diskrete og kontinuerte variabler. **indsæt figur** Disse beskrives i et efterfølgende afsnit.

Variabler, der tilhører intervalkategorien, er defineret som variabler, hvor det er muligt at bestemme forskellen mellem to mængder eller to tal. Det kunne eksempelvis være forskellen på temperaturen indenfor, 20°C , og udenfor, 10°C .

Variabler, der tilhører ratiokategorien, er defineret som variabler, hvor det er muligt at bestemme forholdet mellem to mængder eller to tal. Det betyder, at der mellem to tal skal være et særligt forhold. Som eksempel på værdier, der tilhører ratiokategorien, kan være antal hår på hovedet, fart, temperatur målt i kelvin eller tidsforskel. Årsagen til, at temperaturer målt i kelvin, men ikke temperaturer målt i grader celsius, tilhører ratiokategorien er, at værdier i denne kategori skal have et naturligt nulpunkt.

2.2.1 Diskrete variabler

En underkategori til kvantitative variabler er diskrete variabler. En variabel er diskret, hvis man kan tælle værdierne i den, som eksempelvis øjne på en terning. Enhver variabel der kan beskrives, som “antallet af...” er en diskret variabel. Værdierne i en diskret variabel er heltal, eksempelvis kan det ikke siges, at en person har 2,5 søskende.

Nominale og ordinale variabler er diskrete variabler. Det kan dog være forvirrende at kende forskel på diskrete og kontinuerte variabler, da man ofte afrunder kontinuerte variabler til heltal når man måler dem, så de bliver diskrete. Alder er ét eksempel på en afrunding af et kontinuert tal. Eksempelvis siger man typisk 21 år, når man faktisk er mellem 21 og 22 år. Grunden til, at det er vigtigt at

genkende om variablen er kontinuert eller diskret, er at den viden bruges til at bestemme, hvilken statistisk metode, der skal benyttes til dataanalysen. (Kilde: (?))

2.2.2 Kontinuerte variable

Hvis en variabels værdier tilhører en overtællelig mængde, siges den at være kontinuert. Et eksempel på en kontinuert variabel, kan være højden på et tilfældigt menneske i en population. Her kan højden have et antal decimaler, der går mod uendeligt, for eksempel $195,00000003\text{cm}$.

For kontinuerte variable bruger man tæthedsfunktionen, $f(x)$. Her er arealet, hvilket svarer til 1 altså 100 procent, under tæthedsfunktionen sandsynligheden for, at en stokastisk variabel, X , befinder sig i intervallet (a, b) , givet ved,

$$P(a < X < b) = \int_a^b f(x)dx$$

Den typiske kontinuerte fordeling er en normalfordeling. Dels fordi den fremgår i mange målinger i virkeligheden, og dels fordi, den centrale grænseværdisætning siger, at ved et højere antal observationer, vil alle stokastiske variables gennemsnitsfordeling antage en normalfordeling.

2.3 Stokastiske variable

2.4 Outliers

En outlier kan beskrives som en observation, der afviger betydeligt fra resten af den stikprøve, observationen stammer fra (WIKI). Denne beskrivelse uddybes nedenfor, hvor to forskellige metoder til at identificere outliers gennemgås. Afsnittet er skrevet med udgangspunkt i (? , s. 48-55).

Den første metode til at identificere outliers tager udgangspunkt i fordelingen af observationerne. Hvis fordelingen er tilnærmelsesvis klokkeformet, gælder figur, hvor μ er middelværdien og σ er standardafvigelse.

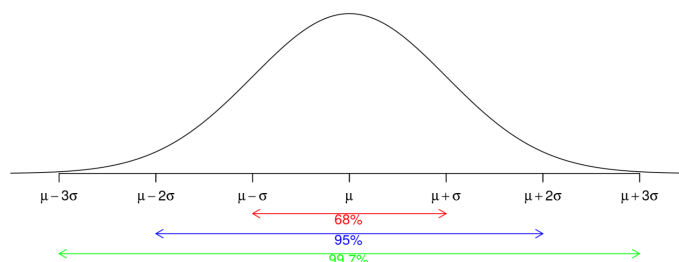


Figure 2.1: Figurtekst

Det gælder altså, at $\approx 68\%$ af observationerne findes inden for én standardafvigelse fra middelværdien, $\approx 95\%$ inden for to standardafvigelser og $\approx 99,7\%$ inden for tre standardafvigelser.

På baggrund af dette, kan en outlier defineres som en værdi, der ligger et vist antal standardafvigelser væk fra middelværdien. Som udgangspunkt siges det, at hvis en værdi er mere end 3 standardafvigelser væk fra middelværdien, kan den ses som en outlier.

En anden måde at beskrive fordelingen af en observation er ved hjælp af percentiler - et punkt for hvilket en hvis mængde af data antager værdier, der ligger over eller under punktet. Her en medianen et specialtilfælde af en percentil, 50-percentilen, hvor halvdelen af data antager værdier under punktet, og den anden halvdel over punktet. Generelt defineres en percentil som,

Definition 2.1. p-percentilen er et punkt, således at $p\%$ af observationerne antager værdier under punktet, og $(100 - p)\%$ af observationerne antager værdier over punktet.

To andre særlige tilfælde af percentiler er 25-percentilen, den nedre kvartil (Q_L), og 75-percentilen, den øvre kvartil (Q_U). Ved hjælp af Q_L og Q_U kan variabiliteten af observationerne beskrives som afstanden mellem de to kvartiler. Denne afstand benævnes *Inner Quartile Range (IQR)*, og beregnes som $IQR = Q_U - Q_L$.

Når både minimum, nedre kvartil, median, øvre kvartil og maksimum er fundet, kan disse benyttes til at tegne et boksplot af observationerne. Et boksplot giver et grafisk overblik af midten og variabiliteten af observationerne. Et boksplot ser ud, som på figuren nedenfor.

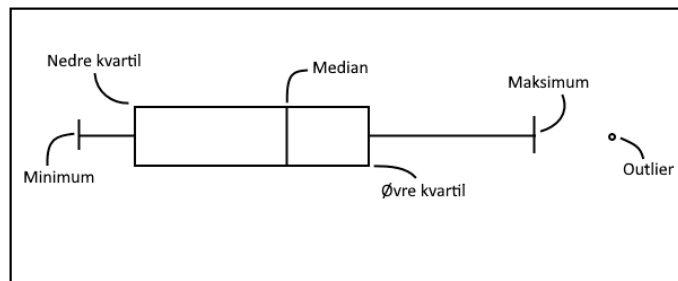


Figure 2.2: Figurtekst

På baggrund af *IQR* og boksplottet kan outliers nu defineres som data, der antager værdier mere end $1,5(IQR)$ gange over Q_U eller under Q_L .

Outliers er altså værdier, der er ekstreme nok til at påvirke en dataanalyse i forkert retning. Derfor benyttes forskellige metoder til at identificere sådanne

outliers, som derefter kan fjernes i en rengøring af datasættet.

2.5 Problemformulering

Hvorvidt kan simulationsstudier bidrage til at højne kvaliteten af dataanalyser?

1. Hvilke faldgruber skal man være opmærksom på?
2. Hvordan simulerer man data?
 - a. Hvordan sikrer man en høj kvalitet af simuleret data?
3. Hvad kan man inferere ud fra simuleret data?
4. Hvilke alternativer finder der til statistisk inferens? (Bootstrapping)

Chapter 3

Faldgruber i en dataanalyse

Der er flere forskellige forhold at tage hensyn til, inden en dataanalyse påbegyndes. Nogle eksempler på disse forhold gennemgås i dette kapitel, herunder variabeltyper, estimatorens egenskaber, fordelinger og antallet af kovariater. Afsnittet er primært skrevet med inspiration fra (?) og (?).

3.1 Variabeltyper

I dette afsnit gennemgås... fordi...

Som nævnt i **AFSNITTET**, findes der forskellige typer af variable. Dette er en vigtig faktor at tage hensyn til, da det har indflydelse på, hvilke matematiske operationer, der kan udføres på variablen. Desuden har variabeltypen indflydelse på, hvilken type diagram de kan illustreres i. En kort opsummering af, hvilke operationer, der kan udføres på hvilke variable, ses på tabellen nedenfor.

Table 3.1: En oversigt over de forskellige variable, og hvilke typer af matematiske operationer, der kan udføres på dem.

Kvalitative variable	Kvantitative variable		
Nominal	Ordinal	Interval	Ratio
Andel	Median	Middelværdi	Middelværdi
Frekvens	Kvartiler	Standardafvigelse	Standardafvigelse

Generelt gælder det, at hvis en variabel tillader en bestemt operation, tillades samme operation på variable, der står til højre i tabellen.

En anden væsentlig, omend ikke lige så matematisk faldgrube, er symboler. Der defineres nøjagtigt, hvad en parameter, et estimat og en estimator er, og de hver

et specifikt symbol. Generelt gælder, at græske bogstaver definerer en parameter, det tilsvarende latinske bogstav et estimat, og et stort latinsk bogstav en estimator. I tabellen nedenfor er det opsummeret, hvad fire forskellige parametre er givet som, og hvad det tilsvarende er i stikprøven og estimatoren.

Table 3.2: En oversigt over de forskellige symboler, som angiver henholdsvis en parameter, et estimat og en estimator.

Population (Parametre)	Stikprøve (Estimator)	Estimatorer
my	my_hat	My_hat
sigma	s	S
sigma ²	s ²	S ²
pi	pi_hat	Pi_hat

Det er derfor klart, at der skal være enighed om, hvad en parameter og et estimat er, og hvad forskellen er på et estimat og en estimator, før en dataanalyse kan udføres korrekt.

3.2 Estimatorens egenskaber

I dette afsnit gennemgås de forskellige egenskaber, unbiased, efficient, konsistent og sufficient, en estimator kan have. Disse begreber giver et indblik i, hvor god en estimator er, og derved også, hvor valide resultater, en dataanalyse på baggrund af den estimator, vil bibringe.

En estimator er en stikprøvefunktion, der udfører en beregning på baggrund af værdier i en stikprøve, for at estimere parametrene i en population. Eksempelvis er middelværdi, $mean = \frac{\sum_{\text{observationer}}}{\#_{\text{observationer}}}$, en estimator, der beregner et estimat for en populations middelværdi.

Der findes overordnet set fire forskellige måder at måle, hvorvidt en estimator er passende for ens data. Disse er, om estimatoren er forventningsret, efficient, konsistent og sufficient, som defineret nedenfor.

Definition 3.1. Estimatorers egenskaber (?)

Forventningsret: Den forventede værdi for en forventningsret estimator er lig parameteren $E(\hat{Y}) = \mu$.

Efficient: En efficient estimator er den estimator med lavest varians.

Konsistent: Variansen af en konsistent estimator falder, når størrelsen af stikprøven stiger.

Sufficient: For en sufficient estimator gælder det, at når estimatet er beregnet, skal stikprøven ikke indeholde yderligere informationer om parameteret.

Hver af disse egenskaber har indvirkning på, hvor brugbare resultater, en data-analyse vil bibringe. En estimator, der ikke er forventningsret, vil ikke give et korrekt bud på populationens middelværdi. Ligeledes vil en estimator, der ikke er efficient, have en større varians end mindst en anden estimator, hvilket leder til mindre brugbare resultater. Det er derfor vigtigt, at en estimator undersøges grundigt, og dens egenskaber bestemmes, inden den tages i brug.

3.3 Populationens fordeling

I dette afsnit gennemgås, hvilken indflydelse en populations fordeling har på fordelingen af estimaterne og de tilhørende estimatorer. Dette er nødvendigt, da forskellige regneregler og metoder, der gælder for eksempelvis en normalfordeling ikke nødvendigvis gælder for eksempelvis en uniform fordeling.

- Population højreskæv \rightarrow estimat højreskæv og estimator højreskæv
- Generel foldningsregel
- Den centrale grænseværdisætning

Sætning 3.1. *Den Centrale Grænseværdisætning* (?, s. 93).

Ved tilfældig stikprøveudtagning af tilpas stor størrelse, vil fordelingen af stikprøvegennemsnittet blive tilnærmelsesvist normalfordelt.

Chapter 4

Hvad er kovariater

Kovariater er variable som beskriver responsvariablen. Responsvariablen er den variabel der ændre sig i forhold til de forklarende variable (kovariater). Hvis man for eksempel vil lave en analyse af mænd og kvinders indkomst på arbejdsmarkedet, vil det give mening at tilføje en variabel der måler på arbejdserfaring, da den vil kunne antages at have en indflydelse på indkomsten. Kontrolvariablen der tilføjes som måler arbejdserfaring kaldes for en kovariat. Denne type sammenligning er kaldet “analysis of covariance”. Hvis der ikke blev taget højde for, den ekstra variabel, arbejdserfaring og dermed kun ses på forholdet mellem indkomst og køn, ville den type sammenligning være “analysis of variance”.

4.1 Hvorfor og hvordan bruges det

Fordi at indflydelse kan ændre sig efter at der kontrolleres for variable (kovariater), kan resultaterne for “analysis of covariance” og “analysis of variance” være forskellige. For eksempel med indkomsten mellem mænd og kvinder, kan det siges at mere arbejdserfaring har positiv indflydelse på indkomsten. Hvis mænd tenderer til at have mere arbejdserfaring end kvinder, vil resultatet ændre sig, afhængig af om man kontrollerer for variabelen arbejdserfaring.

4.2 Tilpas antal kovariater

Chapter 5

Simulering af variabler

Simuleringer er generering af estimater på mulige udfald, og på den måde efterligne virkeligheden ved simulering. Formålet er derved, at generere tilfældigt estimerede værdier ud fra en model, der simulerer virkeligheden, hvilket muliggør yderligere analyser.

En definition på simulation er;

A situation or event that seems real but is not real, used especially in order to help people deal with such situations or events. - Cambridge Dictionary [<https://dictionary.cambridge.org/dictionary/english/simulation>]

Ud fra definitionen, er formålet altså ved simuleringer at efterligne virkeligheden, så de analyser der gøres på baggrund af simuleringerne, kan bruges i virkeligheden når lignende situationer opstår. Brancher hvor simulationer er et yderst vigtigt redskab, er i motorsporten. I Formula 1, benytter holdene sig af simulatore, hvor de genskaber bilerne og derved kan teste nye dele af inden de behøver at producere dem i virkeligheden for at spare penge, så de derved kan analysere, hvad der for eksempel producere mere *down-force* af forskellige dele, og så etablere dem i stedet for, at skulle teste dem i virkeligheden. [<https://careers.mercedesamgf1.com/facilities/simulator/>]

5.1 Simulering i R

Simuleringer i R gøres ved at estimere udfald fra en fordeling, hvor der bliver genereret "tilfældige" tal. Eftersom tilfældige tal står i citationstegn, er det fordi, at der refereres til pseudo-tilfældige tal. Disse generede tal forekommer at være tilfældige, men er det reelt set ikke helt tilfældige. [https://en.wikipedia.org/wiki/Pseudorandom_number_generator]

Tallene er genereret ud fra et *seed*, som i R bestemmes ved `set.seed("værdi")`. Generere man tal ud fra samme *seed*, så vil værdierne altså være identiske.

Fordelinger som der blandt andet kan simuleres i R er normalfordelinger, binomialfordelinger og uniformfordelinger.

En normalfordeling kan simuleres på følgende måde:

```
set.seed(1)
rnorm(10, mean = 0, sd = 1)

## [1] -0.6264538  0.1836433 -0.8356286  1.5952808  0.3295078 -0.8204684
## [7]  0.4874291  0.7383247  0.5757814 -0.3053884
```

Først sættes et *seed*, så man kan rekonstruere samme simulation igen. Næste linje startes med at skrive "rnorm" hvor "r" står for tilfældigt genererede tal, og "norm" for en normalfordeling. Inde i parantesene angives antallet af værdier der skal genereres, som bliver genereret ud fra en middelværdi på 0 og en standardafvigelse på 1. Eftersom middelværdien er 0 og standardafvigelsen er 1, kaldes denne normalfordeling for en standard normalfordeling eller Z-fordeling.

Ligeledes kan man simulere binomialfordelinger og uniformfordelinger:

```
set.seed(1)
rbinom(10, size = 2, prob = 0.5)
```

```
## [1] 1 1 1 2 0 2 2 1 1 0
```

Som før sættes et *seed*, og der skrives "r" før fordelingen der simuleres. Ligeledes er første værdi antallet af værdier der skal genereres. I binomialfordelingen er der angivet *size*, som er antallet success/faliure (1/0) repetationer, hvor hvor *prob* er sandsynligheden for success.

```
set.seed(1)
runif(10, min = 1, max = 2)
```

```
## [1] 1.265509 1.372124 1.572853 1.908208 1.201682 1.898390 1.944675 1.660798
## [9] 1.629114 1.061786
```

Uniformfordelingen er derimod angivet med en minimum og maximum-værdi, hvor der genereres værdier imellem.

5.1.1 Sample og replicate (Bootstrapping)

Derudover kan der simuleres på baggrund af observeret data. Ved at benytte *sample* og *replicate* funktionerne kan der dannes nye simulationer. *Sample* tager en stikprøve af den observerede data, hvor *replicate* kan gentage forskellige stikprøver.

```
set.seed(1)
Z_fordeling <- rnorm(1000, mean = 0, sd = 1)
```

```
mean(Z_fordeling)
```

```
## [1] -0.01164814
```

Her eksekveres en Z-fordeling af 1000 observationer hvor middelværdien, -0.01164814, printes. Denne middelværdi vil variere afhængigt af det *seed* der benyttes.

Dernæst kan der foretages en sample af Z_fordeling.

```
set.seed(1)
```

```
Z_sample <- sample(Z_fordeling, size = 10, replace = TRUE)
```

```
Z_sample
```

```
## [1] 1.1654620 0.2109073 -0.6816605 0.3439100 1.1594245 -0.8132443
```

```
## [7] -0.0505657 -0.9398293 0.5050676 1.4645873
```

```
mean(Z_sample)
```

```
## [1] 0.2364059
```

Her fremgår der 10 værdier, som er taget fra Z_fordeling, som gemmes i Z_sample. *Replace* gør så de værdier der tages og gemmes i Z_sample bliver lagt tilbage, og kan derved blive taget igen, så man kan få den samme værdi flere gange.

Derefter er det muligt at gentage disse stikprøver ved brug af *replicate*

```
set.seed(1)
```

```
Z_replicate <- replicate(100, {  
  x <- mean(sample(Z_fordeling, size = 10, replace = TRUE))  
})
```

```
Z_replicate
```

```
## [1] 0.025838777 -0.203406868 -0.018837976 0.124819713 -0.164830739
```

```
## [6] 0.333733810 -0.423105712 -0.184935071 0.032046950 -0.166711138
```

```
## [11] 0.128328946 -0.003149890 0.120181675 0.114612739 0.196252732
```

```
## [16] -0.503558064 -0.057580606 0.448834352 0.302573714 0.349289426
```

```
## [21] 0.042763308 -0.449645007 -0.567764002 -0.246353993 0.241230288
```

```
## [26] 0.118362694 0.360972297 -0.502740430 -0.294227992 0.116454913
```

```
## [31] -0.458241385 0.145491729 0.468018395 -0.336889894 -0.033331555
```

```
## [36] 0.003202705 -0.230521260 0.002509074 -0.104908678 0.577232470
```

```
## [41] 0.405361935 0.571805412 -0.242967988 -0.070361799 -0.204749550
```

```
## [46] 0.529131107 0.123788480 0.428641140 0.421487366 0.296227471
```

```
## [51] -0.021464536 -0.563489156 -0.096866861 0.011460957 0.490675566
```

```
## [56] 0.118853950 -0.048831160 -0.433364372 -0.158288527 0.228563345
```

```
## [61] 0.383487805 -0.348257409 0.251253782 0.071396170 0.263839481
```

```
## [66] -0.659980071 -0.272080635 -0.065224329 0.133837717 -0.278633082
```

```
## [71] -0.543857905  0.066031905  0.675680619  0.172629193 -0.132092862
## [76]  0.269779598  0.150154779 -0.214823863 -0.098373108 -0.102126967
## [81] -0.350372760  0.502067558  0.111435486 -0.269840347  0.159325611
## [86]  0.033092712 -0.180166957  0.284144813 -0.108679711  0.028228077
## [91] -0.375649589 -0.224793309 -0.128662553  0.074568112  0.052000689
## [96] -0.061919916  0.245951870  0.073717038 -0.297133939  0.034445758
```

```
mean(Z_replicate)
```

```
## [1] 0.004120227
```

Her bliver en stikprøve af Z fordeling foretaget 10x100 gange, hvorefter middelværdien på de 100 gentagelser bliver fundet.

Det med at tage en et stort antal stikprøver af nogle observationer er også kaldet bootstrapping, hvilket forklares i [AFSNIT XXXX]

Henvises der ikke til noget er kilden: [https://bookdown.org/rdpeng/rprogdatascience/simulation.html] **AFSNIT SLUT**

5.1.2 Simulering af nominale variabler

En nominal variabel kan for eksempel være køn, en måde at simulere dette på kan være ved en binomialfordeling.

```
set.seed(1)
gender <- rbinom(100, size = 1, prob = 0.5)
mean(gender)
```

```
## [1] 0.48
```

Hvis kvinde er en success og mand er en failure, så er der derved en forekomst på 48% kvinder og 52% mænd. Derved har binomialfordelingen taget 100 værdier, en gang, hvor der har været 50% sandsynlighed for success, som har været kvinde. Da middelværdien er 0.48, betyder det så at der ud fra 100 observationer er 48 kvinder.

5.1.3 Simulering af ordinale variabler

En ordinal variabel, som der kan simuleres kan være minutter det tager at løse en opgave, hvilket kan simuleres ved en uniformfordeling.

```
library(mosaic)
```

```
set.seed(1)
opgave <- as.integer(runif(100, min = 1, max = 10+1))
gf_histogram(~opgave, bins = 10)
```