Forelesning 1 document

Eirik Eriksen Heen

```
##### Start up #####
rm(list = ls()) # Empties all data

options(scipen=10) # writes 10 scipens before scientific script
options(digits=10) # writes up to 10 digits


# loading packages
library(tidyverse)
```

```
Warning: package 'ggplot2' was built under R version 4.3.3
```

```
-- Attaching core tidyverse packages ------------------------ tidyverse 2.0.0 --
v dplyr     1.1.4     v readr     2.1.5
v forcats   1.0.0     v stringr   1.5.1
v ggplot2   3.5.1     v tibble    3.2.1
v lubridate 1.9.3     v tidyr     1.3.1
v purrr     1.0.2
-- Conflicts ------------------------------------------ tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
i Use the conflicted package (<http://conflicted.r-lib.org/>) to force all conflicts to beco
```

```
library(NHANES)
library(gt)
```

```
Warning: package 'gt' was built under R version 4.3.3
```

```
# Kode for a kunne bruke norske bokstaver
Sys.setlocale(locale="no_NO")
```

```
Warning in Sys.setlocale(locale = "no_NO"): using locale code page other than
65001 ("UTF-8") may cause problems
```

```
[1] "LC_COLLATE=no_NO;LC_CTYPE=no_NO;LC_MONETARY=no_NO;LC_NUMERIC=C;LC_TIME=no_NO"
```

```
NHANES <- NHANES
```

Velkommen til Sok-2009

I dette emnet skal dere lære om statistikk. Dette inkluderer:

- Deskriptiv statistisk
- Sannsynlighets regning
- Fordelinger
- Interferens
- Regresjon

Dette emnet består av flere deler:

- Forelesninger
- Seminarer
- Datacamp
- Arbeidskrav
- Eksamen

Forelesninger

- Vi blir å starte intensivt før det blir for mange innleveringer og arbeidskrav i *Den nordiske modellen*.
- I forelesningene blir jeg å gå igjennom teori, kode og intuisjon.
- Vi blir å følge delvis oppbyggingen til Datacamp modulene og pensumbok.

Innleveringer

I dette emnet blir det 2 innleveringer:

- Datacamp: Alle modulene er obligatorisk og må gjøres ferdig innen 08.11 klokken 14:00.
- Muntelig presantasjon av prosjekt oppgave.

Eksamen

I dette emnet blir det 3 mappe innleveringer:

- **Mappe oppgave 1:** Grunnlengene statistikk (25%)
- **Mappe oppgave 2:** Regresjon (25%)
- **Mappe oppgave 3:** Prosjekt oppgave (50%)

Utvalg og populasjon

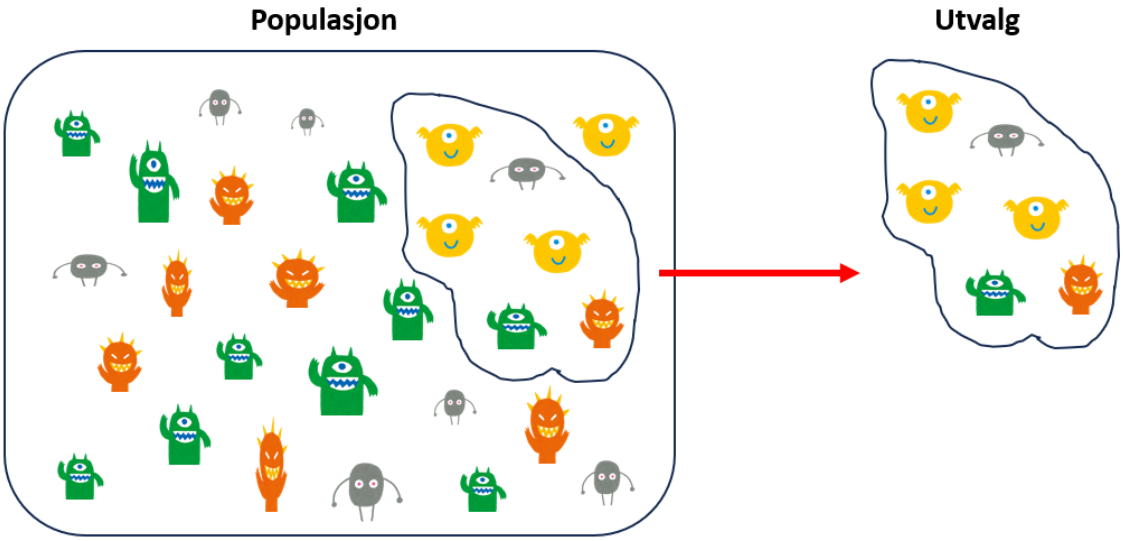
Når vi jobber med statistikk er det viktig å skille de to gruppene, utvalg og populasjon.

Populasjon:

- Populasjonen er den fullstendige mengden av individer eller enheter vi ønsker å studere.
- Eksempel: Alle innbyggerne i et land, alle studentene ved et universitet, eller alle bedriftene i en bestemt bransje.
- Populasjonen inneholder alle mulige data vi kan samle om et fenomen.
- Viktig!: en populasjon trenger ikke være alle som bor i et land, men gruppen vi er interesert i å undersøke.

Utvalg:

- Et utvalg er en mindre gruppe som er trukket fra populasjonen, som brukes til å gjøre slutninger om hele populasjonen.
- Eksempel: En undersøkelse blant 1000 personer valgt fra et lands befolkning.
- Utvalget brukes fordi det ofte er upraktisk eller umulig å samle data fra hele populasjonen.



- **Deskriptiv statistikk:**
  - Forklarer utvalget
- **Inferensiell statistikk:**
  - Prøver å si noe om utvalget

Hva er statistikk?

Statistikk er vitenskapen om innsamling, analyse, tolkning, presentasjon og organisering av data. Det gir oss verktøy for å forstå og arbeide med data på en strukturert måte, og for å trekke konklusjoner basert på data.

- **To hovedområder innen statistikk:**
  - **Deskriptiv statistikk:**
    - Beskriver og oppsummerer data
    - Eksempler inkluderer: gjennomsnitt, median, standardavvik, diagrammer
  - **Inferensiell statistikk:**
    - Bruker et utvalg av data for å trekke konklusjoner om en populasjon
    - Inkluderer: hypotesetesting, konfidensintervall, prediksjon
- **Anvendelser av statistikk:**

  - Økonomi
  - Medisin
  - Psykologi
  - Samfunnsvitenskap
- **Hvorfor bruker vi statistikk:**
  - For å bryte ned informasjon til en mer forståelig versjon
  - For å forstå mønstre og ta informerte beslutninger basert på data

Hva kan statistikk gjøre og hva kan det ikke?

Statistikk er et kraftig verktøy for å analysere data og trekke konklusjoner, men det har også sine begrensninger. Det er viktig å forstå hva statistikk kan og ikke kan gjøre for å bruke det effektivt og unngå misforståelser.

Hva statistikk kan gjøre:

- **Beskrivelse:** Statistikk kan hjelpe oss med å beskrive og oppsummere data. For eksempel, gjennomsnitt, median, modus, varians og standardavvik.
- **Analyse:** Statistikk gir metoder for å analysere data, identifisere mønstre og sammenhenger, som korrelasjon og regresjon.
- **Prognoser:** Statistikk kan brukes til å lage prognoser basert på historiske data. Eksempler inkluderer tidseriemodeller og prediktive analyser.
- **Testing:** Statistiske metoder kan brukes til å teste hypoteser og trekke konklusjoner om en populasjon basert på et utvalg.

Hva statistikk ikke kan gjøre:

- **Forutsi nøyaktige resultater:** Statistikk kan bare gi sannsynligheter og ikke definitive resultater.
- **Eliminere usikkerhet:** Selv om statistikk kan redusere usikkerhet, kan det aldri eliminere det fullstendig.
- **Erstatte ekspertise:** Statistiske resultater må tolkes i kontekst av faglig ekspertise og erfaring.

Hvorfor er statistikk viktig for en økonom?

- **Dataanalyse:** Økonomer bruker statistikk for å analysere økonomiske data, som BNP, arbeidsledighet og inflasjon.
- **Beslutningstaking:** Statistikk hjelper økonomer med å ta informerte beslutninger basert på data.
- **Politikkutforming:** Statistiske analyser brukes til å vurdere effekten av økonomiske politikktiltak.
- **Økonometriske modeller:** Økonomer bruker statistiske metoder for å utvikle modeller som beskriver økonomiske fenomener.

Hva er en variabel? Og hva er en observasjon?

For å kunne analysere data, må vi forstå de grunnleggende elementene i datasett: variabler og observasjoner. En variabel representerer en egenskap eller et kjennetegn som kan måles, mens en observasjon er et enkeltdatapunkt i datasettet.

Variabel

En variabel er en egenskap eller et kjennetegn som kan måles og som kan anta forskjellige verdier.

Eksempler: alder, inntekt, kjønn, pris på en vare, etc.

Typer variabler:

- Kategoriske (nominale, ordinal)
- Numeriske (diskrete, kontinuerlige)

Observasjon

En observasjon er en enkeltdatainnsamling for en bestemt variabel eller sett av variabler.

Eksempel: inntekten til en spesifikk person i en undersøkelse er en observasjon.

Eksemple på et datasett

```r
# Opprette et eksempeldataframe i R
df <- data.frame(
  Navn = c("Ada", "Bent", "Charlie", "David"),
  Alder = c(25, 30, 35, 40),
  Høyde_cm = c(162, 175, 168, 180),
  Øyefarge = c("Blå", "Brun", "Grønn", "Brun"),
  Karakter = factor(c("B", "A", "C", "E"), levels=c("A","B","C","D","E","F"))
)

# Skriv ut dataframe ved å bruke gt() for å gjøre tabellen penere
df %>%
  gt()
```

Navn	Alder	Høyde_cm	Øyefarge	Karakter
Ada	25	162	Blå	B
Bent	30	175	Brun	A
Charlie	35	168	Grønn	C
David	40	180	Brun	E

I tabellen over er radene observasjoner og kolonene er variabler.

Hvordan kan vi måle en variabel?

For å analysere data nøyaktig, må vi forstå hvordan variabler kan måles. Måten vi måler en variabel på, påvirker hvilke analyser som kan utføres og hvilke konklusjoner som kan trekkes. Når vi tenker på datasettet vi lagde tidligere, er høyde greit å forstå siden den er numerisk, og øyefarge er en kategori. Men hva med karakterer?

Måter å måle variabler på:

1. **Nominalnivå:**
    - Variabler som kategoriserer data uten noen naturlig rekkefølge.
    - Eksempler: kjønn, farge, nasjonalitet.
2. **Ordinalnivå:**
    - Variabler som har en naturlig rekkefølge, men avstandene mellom verdiene er ikke nødvendigvis like.

- Eksempler: utdanningsnivå, tilfredshetsgrader, karakterer.

3. **Intervallnivå:**
   - Variabler som har like intervaller mellom verdiene, men ingen sann nullpunkt.
   - Eksempler: temperatur i Celsius, kalenderår.
4. **Forholdstall:**
   - Variabler som har like intervaller mellom verdiene og et meningsfylt nullpunkt.
   - Eksempler: alder, inntekt, vekt.

```
str(df)
```

```
'data.frame':   4 obs. of  5 variables:
 $ Navn    : chr  "Ada" "Bent" "Charlie" "David"
 $ Alder   : num  25 30 35 40
 $ Høyde_cm: num  162 175 168 180
 $ Øyefarge: chr  "Blå" "Brun" "Grønn" "Brun"
 $ Karakter: Factor w/ 6 levels "A","B","C","D",..: 2 1 3 5
```

```
df %>%
  gt()
```

Navn	Alder	Høyde_cm	Øyefarge	Karakter
Ada	25	162	Blå	B
Bent	30	175	Brun	A
Charlie	35	168	Grønn	C
David	40	180	Brun	E

I tabellen over er høyde et eksempel på en numerisk variabel (forholdstall), øyefarge er en kategorisk variabel (nominalnivå), og karakterer er en ordinal variabel. Forståelse av disse målenivåene er essensiell for å kunne velge riktig statistisk metode og tolke resultatene riktig.

Disse kalles ofte nivåene av måling. Nominalnivå er det "laveste" nivået, fordi det gir oss den groveste inndelingen. Neste er ordinalnivå, som gir litt mer informasjon, etterfulgt av intervallnivå, og til slutt forholdstallsnivå som anses som det "høyeste". Dette hierarkiet er viktig for å bestemme hvilke statistiske tester og analyser som er passende.

Nominalnivå

Nominalnivå er det mest grunnleggende nivået av måling. Det brukes til å kategorisere data uten noen form for rangering eller orden. Variabler på dette nivået kan bare klassifiseres og telles, de kan ikke rangordnes eller måles på noen meningsfull måte. Kategoriene er gjensidig utelukkende, noe som betyr at en observasjon kun kan tilhøre én kategori om gangen.

Egenskaper ved nominalnivå

- Ingen naturlig rekkefølge: Kategoriene har ingen innebygd ranger eller hierarki.
- Kun kategorisering: data klassifiserer i ulike grupper eller kategorier.
- Ingen numerisk betydning: Tall som brukes for å representere kategorier har ingen kvantitativ verdi.
- Gjensidig utelukkende: Hver observasjon tilhører kun én kategori.

Eksempler på nominalnivå:

- Kjønn (mann, kvinne, ikke binær, transpersoner)
- Øyenfarge (blå, brun, grønn)
- Merke på biler (Toyota, Ford, BMW)
- Type kjøretøy (bil, motorsykkel, sykkel, buss)

Bruk i statistikk:

- **Frekvensfordelinger:** Telle antall observasjoner i hver kategori.
- **Modus:** Identifisere den mest vanlige kategorien.

I bildet nedenfor ser vi forskjellige typer transportmidler. Hvis spørsmålet er hvilke typer transportmidler dataene her inneholder, kan vi begynne å gruppere dataene. For eksempel kan vi si at det er: 4 bilder, 2 sykler, 2 tog, 1 buss og 1 gravemaskin.

Det finnes ingen måte å rangere disse typene på, og det er ingen overlapp mellom kategoriene; en sykkel er ikke et tog, og så videre.

I datasett kodes nominalnivå hovedsakelig som karakter eller faktor (uten nivåer), men det kan også kodes med tall hvor et tall tilsvarer en kategori.

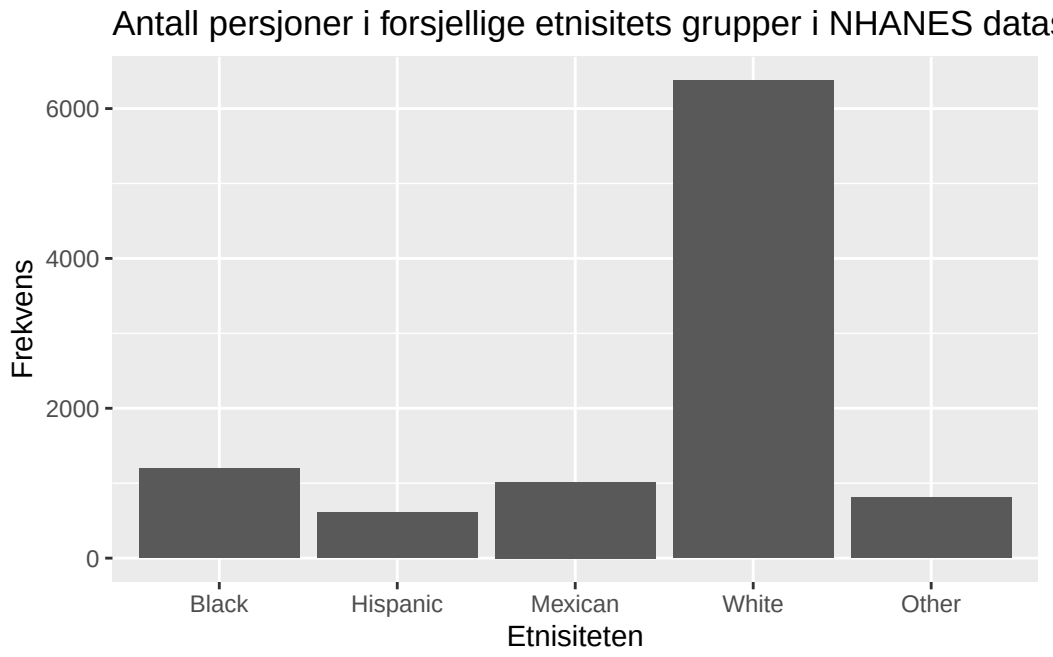
La oss se på et eksempel med etnisitet i datasettet NHANES.

```
# Kaller opp datasettet NHANES
NHANES %>%
  # Grupperer på den variablen vi er interesert i
  group_by(Race1) %>%
  # Teller antallet er hver gruppe som er laget over.
  summarise(Frekvens = n() ) %>%
  # Denne funksjonen gjør tabellen litt penere (vi trenger ikke bruke denne koden)
  gt()
```

Race1	Frekvens
Black	1197
Hispanic	610
Mexican	1015

White	6372
Other	806

```
# Oppretter et ggplot-objekt hvor data kilden defineres og x aksen defineres
ggplot(NHANES, aes(x=Race1)) +
  # Legger til stolper (barplot)
  geom_bar() +
  # Lager en tittle til plottet
  ggtitle("Antall persjoner i forsjellige etnisitets grupper i NHANES datasettet") +
  # Legger til navn på x-aksen
  xlab("Etnisiteten") +
  # Legger til navn på y-aksen
  ylab("Frekvens")
```



Nominalnivå (dummy nivå)

Nominalnivå kan også brukes i sammenheng med dummy variabler, som er binære variabler som tar verdiene 0 eller 1. Dummy variabler brukes ofte i statistisk modellering for å representere kategoriske data som bare kan tilhøre én av to grupper. Brukes ofte for å registrere tilstedeværelse eller fravær av noe.

Egenskaper ved dummy variabler:

- **Binær representasjon:** Verdiene 0 og 1 brukes til å indikere fravær eller tilstedeværelse av en kategori.
- **Enkel kategorisering:** Brukes til å kode kategoriske variabler som bare har to mulige verdier.
- **Praktisk i regresjonsanalyse:** Tillater inkludering av kategoriske variabler i lineære modeller.

Eksempler på dummy variabler:

- **Kjønn:** Kvinne = 0, Mann = 1 (i gamle datasett)
- **Har førerkort:** Nei = 0, Ja = 1
- **Eier bolig:** Nei = 0, Ja = 1

Bruk i statistikk:

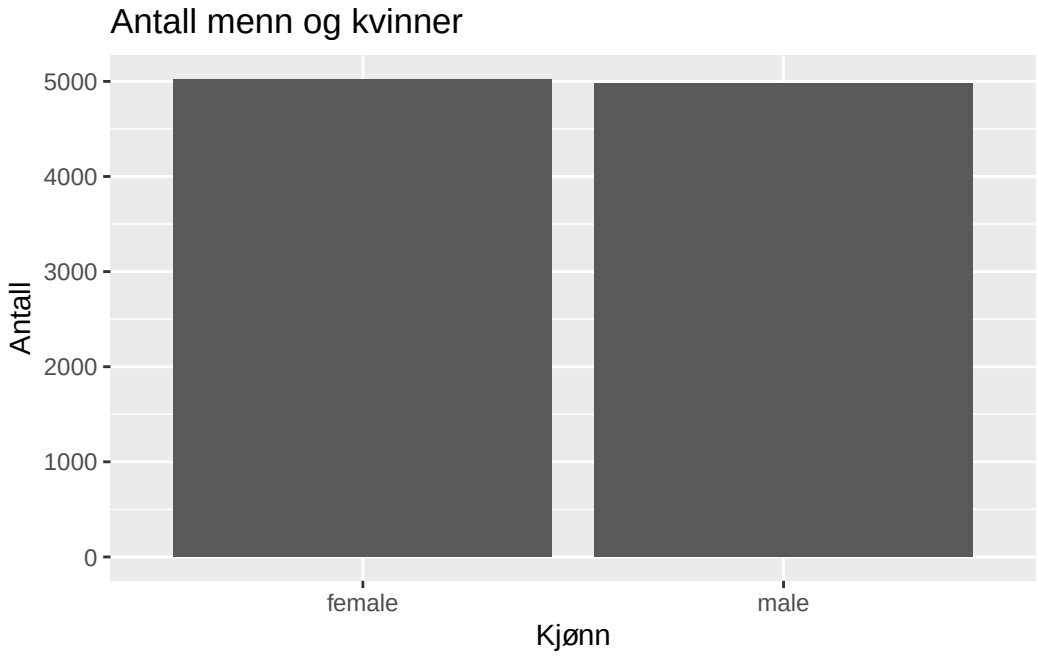
- **Regresjonsanalyse:** Inkludere dummy variabler for å modellere effekten av kategoriske variabler.
- **Tolkning av koeffisienter:** Koeffisientene til dummy variabler representerer forskjellen i den avhengige variabelen mellom de to gruppene.

I NHANES datasettet er kjønn programert som dummy variabel.

```
# Kaller opp datasettet NHANES
NHANES %>%
  # Grupperer etter den interessante variabelen
  group_by(Gender) %>%
  # Teller antallet er hver gruppe som er laget over
  summarise(Frekvens = n()) %>%
  # Formaterer tabellen for en mer ryddig presentasjon
  # Vi trenger ikke bruke denne koden, den er kun for at tabellen skal se litt penere ut. pr
  gt()
```

Gender	Frekvens
female	5020
male	4980

```r
# Oppretter et ggplot-objekt hvor data kilden defineres og x aksen defineres
ggplot(NHANES, aes(x=Gender)) +
  # Legger til stolper (barplot)
  # "fill =" legger til farger for å få litt kontraster
  geom_bar() +
  # Lager en tittle til plottet
  ggtitle("Antall menn og kvinner") +
  # Legger til navn på x-aksen
  xlab("Kjønn") +
  # Legger til navn på y-aksen
  ylab("Antall")
```



Ordianlnivå

Ordinalnivå representerer et høyere nivå av måling enn nominalnivå. I tillegg til å kategorisere data, innebærer ordinalnivå en rangering av kategoriene. Dette nivået gir oss informasjon om rekkefølgen på kategoriene, men ikke nødvendigvis avstandene mellom dem.

Egenskaper ved ordinalnivå:

- **Naturlig rekkefølge:** Kategoriene har en innebygd rangering eller hierarki.

- **Ingen lik avstand:** Avstandene mellom rangeringene er ikke nødvendigvis like.
- **Rangering uten numerisk verdi:** Tallene som brukes til å representere rangeringene har ingen kvantitativ betydning utover rekkefølgen.

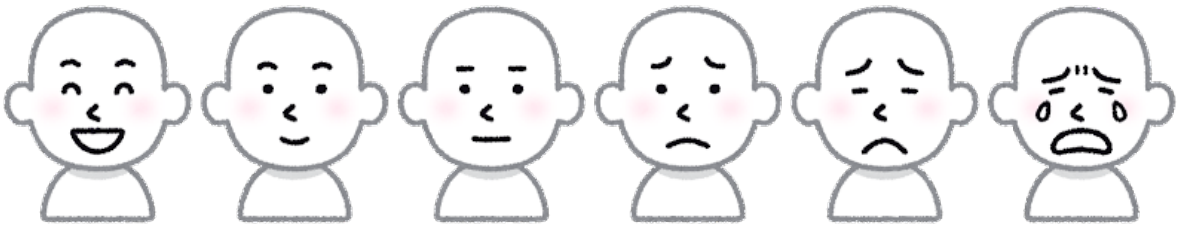
Eksempler på ordinalnivå:

- Utdanningsnivå (grunnskole, videregående, bachelor, master)
- Tilfredshet (svært misfornøyd, misfornøyd, nøytral, fornøyd, svært fornøyd)
- Militær rang (menig, korporal, sersjant, løytnant)

Bruk i statistikk:

- **Median:** Midtpunktet i en rangert liste av data.
- **Rangbaserte tester:** Statistiske tester som Mann-Whitney U og Kruskal-Wallis.

Et eksempel på en ordinal skala.



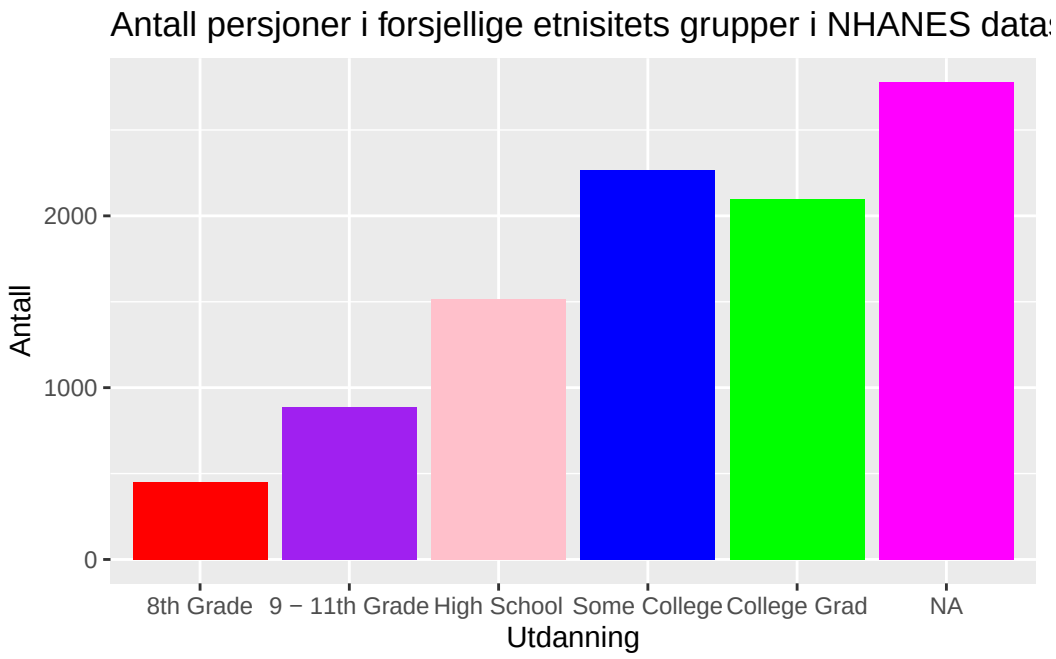
Slik data kan representeres med frekvenstabeller eller stolpediagrammer i R. Ofte vil de være formatert som **factor** med definerte **levels**.

Vi kan se på eksemplet med utdanningsnivå i datasettet NHANES.

```
# Kaller opp datasettet NHANES
NHANES %>%
  # Grupperer etter den interessante variabelen
  group_by(Education) %>%
  # Teller antallet er hver gruppe som er laget over
  summarise(Frekvens = n()) %>%
  # Formaterer tabellen for en mer ryddig presentasjon
  # Vi trenger ikke bruke denne koden, den er kun for at tabellen skal se litt penere ut. pr
  gt()
```

Education	Frekvens
8th Grade	451
9 - 11th Grade	888
High School	1517
Some College	2267
College Grad	2098
NA	2779

```
# Oppretter et ggplot-objekt hvor data kilden defineres og x aksen defineres
ggplot(NHANES, aes(x=Education)) +
  # Legger til stolper (barplot)
  # "fill =" legger til farger for å få litt kontraster
  geom_bar(fill=c("red", "purple", "pink", "blue", "green","magenta")) +
  # Lager en tittle til plottet
  ggtitle("Antall persjoner i forsjellige etnisitets grupper i NHANES datasettet") +
  # Legger til navn på x-aksen
  xlab("Utdanning") +
  # Legger til navn på y-aksen
  ylab("Antall")
```



Intervallnivå

Intervallnivå er et høyere nivå av måling enn både nominal- og ordinalnivå. Dette nivået gir oss informasjon om både rekkefølgen av verdiene og de nøyaktige avstandene mellom dem. Intervallnivå har like intervaller mellom verdiene, men det mangler et naturlig nullpunkt.

Egenskaper ved intervallnivå:

- **Naturlig rekkefølge:** Verdiene har en innebygd rangering.
- **Like intervaller:** Avstandene mellom verdiene er like over hele skalaen.
- **Ingen absolutt nullpunkt:** Nullpunktet er vilkårlig og representerer ikke fraværet av egenskapen som måles.

Eksempler på intervallnivå:

- Temperatur i Celsius eller Fahrenheit (f.eks. 10°C, 20°C, 30°C)
- Kalenderår (f.eks. 1990, 2000, 2010)

Bruk i statistikk:

- **Gjennomsnitt og standardavvik:**
  - Intervallnivådata tillater beregning av gjennomsnitt og standardavvik.
- **Lineær regresjon:**
  - Analyser som forutsetter like intervaller mellom verdier.

Slik data kan representeres med frekvenstabeller eller stolpediagrammer i R. Ofte vil de være formatert som **factor** med definerte **levels**.

Vi kan se på eksemplet med inntektsnivå i datasettet NHANES. Vi presenterer først dataene i en frekvenstabell.

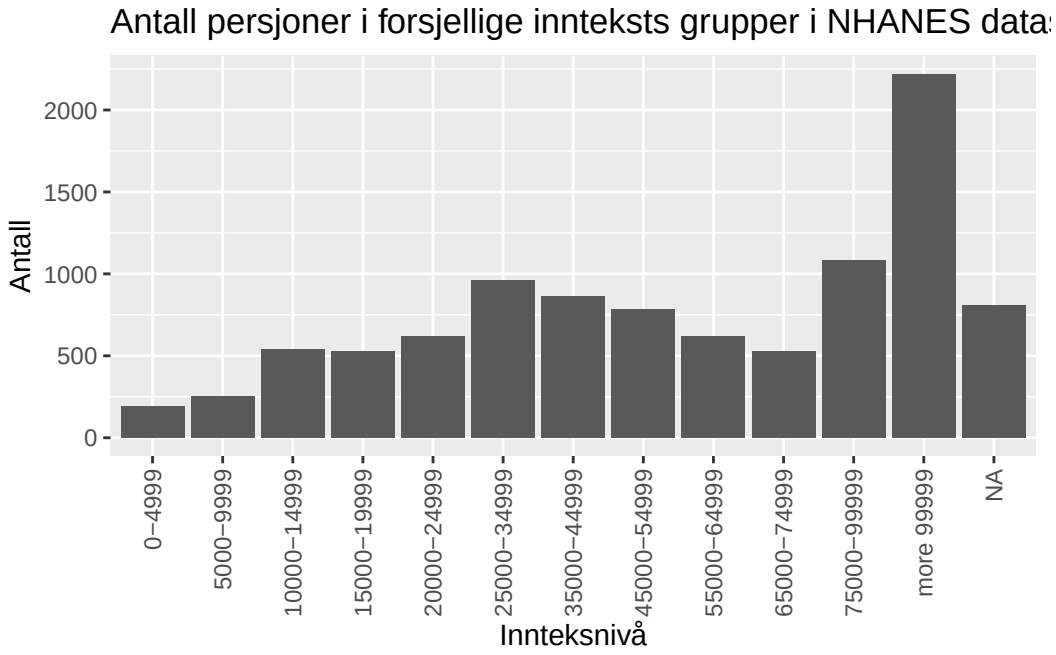
```
# Kaller opp datasettet NHANES
NHANES %>%
  # Grupperer etter den interessante variabelen
  group_by(HHIncome) %>%
  # Teller antallet er hver gruppe som er laget over
  summarise(Frekvens = n()) %>%
  # Formaterer tabellen for en mer ryddig presentasjon
```

```
  # Vi trenger ikke bruke denne koden, den er kun for at tabellen skal se litt penere ut. pr
  gt()
```

HHIncome	Frekvens
0-4999	192
5000-9999	254
10000-14999	543
15000-19999	527
20000-24999	617
25000-34999	958
35000-44999	863
45000-54999	784
55000-64999	621
65000-74999	526
75000-99999	1084
more 99999	2220
NA	811

En tabell gir veldig nøyaktige tall, men igjen kan være vanskelig å få oversikt over dataen, spesielt hvis det er mange kategorier. Et godt alternativ er å lage et stolpediagram.

```
# Oppretter et ggplot-objekt hvor data kilden defineres og x aksen defineres
ggplot(NHANES, aes(x=HHIncome)) +
  # Legger til stolper (barplot)
  geom_bar() +
  # Lager en tittle til plottet
  ggtitle("Antall persjoner i forsjellige innteksts grupper i NHANES datasettet") +
  # Legger til navn på x-aksen
  xlab("Innteksnivå") +
  # Legger til navn på y-aksen
  ylab("Antall") +
  # Denne koden roterer teksten på x-aksen slik at det ikke blir overlapp mellom
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5, hjust=1))
```



Forholdstallsnivå

Forholdstall er det høyeste nivået av måling og gir den mest detaljerte informasjonen. Dette nivået inkluderer både egenskapene til intervallnivå og et absolutt nullpunkt, som betyr at nullverdien representerer fraværet av den målte egenskapen. Dette muliggjør både addisjon, subtraksjon, multiplikasjon og divisjon.

Egenskaper ved forholdstall:

- **Naturlig rekkefølge:** Verdiene har en innebygd rangering.
- **Like intervaller:** Avstandene mellom verdiene er like over hele skalaen.
- **Absolutt nullpunkt:** Nullpunktet representerer fraværet av egenskapen som måles.
- **Meningsfylte forhold:** Mulighet til å sammenligne forhold (f.eks. dobbelt så mye).

Eksempler på forholdstall:

- Alder (f.eks. 20 år, 30 år)
- Inntekt (f.eks. 50 000 kr, 100 000 kr)
- Vekt (f.eks. 60 kg, 80 kg)

- Høyde (f.eks. 160 cm, 180 cm)

Bruk i statistikk:

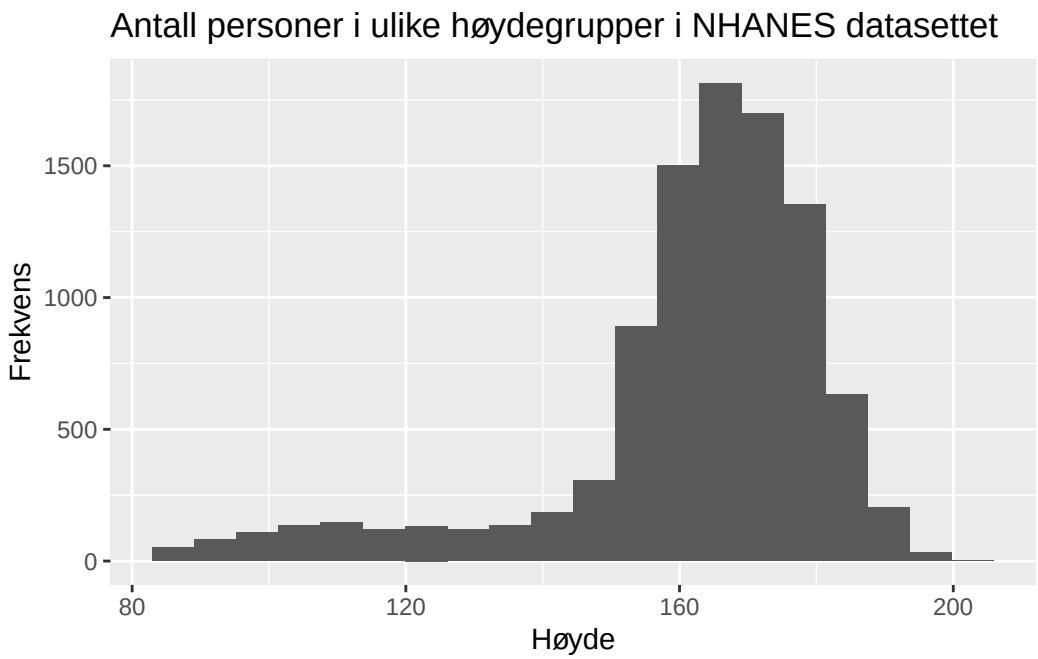
- **Gjennomsnitt og standardavvik:**
  - Forholdstall tillater beregning av gjennomsnitt og standardavvik.
- **Prosentendringer:**
  - Mulighet til å beregne prosentvise endringer og forhold.
- **Regresjonsanalyse:**
  - Analyser som forutsetter forholdstall mellom variabler.

Slik data kan representeres med gjennomsnitt eller histogrammer i R. Ofte vil de være formatert som **intigers** eller **nummeric**.

Histogram omformaterer dataene til å ligne intervalldata. Merk at stolpene er sammenhengende fordi dataene er kontinuerlige, og spenner fra rundt 80 cm til 200 cm:

```
# Initialiserer et ggplot-objekt med definert datakilde og x-akse
ggplot(NHANES, aes(x=Height)) +
  # Tilføyer histogram med spesifisert antall stolper
  geom_histogram(bins = 20) +
  # Setter tittel på plottet
  ggtitle("Antall personer i ulike høydegrupper i NHANES datasettet") +
  # Navngir x-aksen
  xlab("Høyde") +
  # Navngir y-aksen
  ylab("Frekvens")
```

```
Warning: Removed 353 rows containing non-finite outside the scale range
(`stat_bin()`).
```



Histogramet hjelper til å vise fordelingen i datasettet, vi kan se at de fleste ligger rundt høyden 170 cm.

Merk: På engelsk refereres forholdstall til som "ratios".

Type datasett

Når vi arbeider med data, er det viktig å forstå de ulike typene datasett vi kan møte på. Forskjellige datasett gir ulike typer informasjon og krever forskjellige metoder for analyse.

Hovedtyper datasett:

1. **Tverrsnittsdata:**
    - Data samlet inn på ett tidspunkt eller over en kort periode.
    - Representerer et "øyeblikksbilde" av populasjonen.
    - Eksempler: Inntekt og utdanning for en gruppe mennesker samlet inn i 2024, antall salg i en butikk på en bestemt dag.
2. **Tidsseriedata:**
    - Data samlet inn over ulike tidspunkter for å observere endringer over tid.

- Hver observasjon er knyttet til et bestemt tidspunkt.
- Eksempler: Månedlig arbeidsledighet over flere år, daglige aksjekurser.

3. **Paneldata:**
   - Kombinasjon av tverrsnitts- og tidsseriedata.
   - Observasjoner av flere enheter (individer, bedrifter, land) over flere tidspunkter.
   - Eksempler: Inntekt og utdanning for samme gruppe mennesker over flere år, månedlig salg for flere butikker over flere år.
4. **Eksperimentell og ikke-eksperimentell data:**
   - **Eksperimentell data:** Data samlet inn under kontrollerte forhold, hvor forskeren manipulerer variabler for å observere effektene.
   - **Ikke-eksperimentell data:** Data samlet inn uten manipulasjon av variabler, ofte brukt til observasjonsstudier.
   - Eksempler på eksperimentell data: Effekten av et nytt legemiddel i en klinisk studie.
   - Eksempler på ikke-eksperimentell data: Observasjon av kjøpsatferd i en butikk.

Viktigheten av å forstå type datasett:

- **Valg av analysemetoder:**
  - Ulike typer datasett krever forskjellige statistiske metoder.
- **Tolkning av resultater:**
  - Forståelse av datasettets struktur hjelper med riktig tolkning av analysene.
- **Planlegging av datainnsamling:**
  - Kunnskap om type datasett hjelper i design av studier og eksperimenter.