

Case 3

19

Instruksjoner

Denne oppgaven skal løses interaktivt i RStudio ved å legge inn egen kode og kommentarer. Det ferdige dokumentet lagres med kandidatnummeret som navn `[kandidatnummer]_SOK1004_C4_H22.qmd` og lastes opp på deres GitHub-side. Hvis du har kandidatnummer 43, så vil filen hete `43_SOK1004_C4_H22.qmd`. Påse at koden kjører og at dere kan eksportere besvarelsen til pdf. Lever så lenken til GitHub-repositoriumet i Canvas.

Bakgrunn, læringsmål

Innovasjon er en kilde til økonomisk vekst. I denne oppgaven skal vi se undersøke hva som kjennetegner bedriftene som bruker ressurser på forskning og utvikling (FoU). Dere vil undersøke FoU-kostnader i bedriftene fordelt på næring, antall ansatte, og utgiftskategori. Gjennom arbeidet vil dere repetere på innhold fra tidligere oppgaver og øve på å presentere fordelinger av data med flere nivå av kategoriske egenskaper.

Last inn pakker

```
# output | false
rm(list=ls())
library(tidyverse)
```

```
-- Attaching packages ----- tidyverse 1.3.2 --
v ggplot2 3.3.6      v purrr   0.3.4
v tibble  3.1.8      v dplyr   1.0.10
v tidyr   1.2.1      v stringr 1.4.1
v readr   2.1.2      v forcats 0.5.2
```

```
-- Conflicts ----- tidyverse_conflicts() --
x dplyr::filter() masks stats::filter()
x dplyr::lag()    masks stats::lag()
```

```
library(rjstat)
```

Attaching package: 'rjstat'

The following object is masked from 'package:dplyr':

id

```
library(gdata)
```

```
Warning in system(cmd, intern = intern, wait = wait | intern,
show.output.on.console = wait, : running command 'C:\Windows\system32\cmd.exe /c
ftype perl' had status 2
```

```
Warning in system(cmd, intern = intern, wait = wait | intern,
show.output.on.console = wait, : running command 'C:\Windows\system32\cmd.exe /c
ftype perl' had status 2
```

gdata: read.xls support for 'XLS' (Excel 97-2004) files ENABLED.

gdata: Unable to load perl libraries needed by read.xls()
gdata: to support 'XLSX' (Excel 2007+) files.

gdata: Run the function 'installXLSXsupport()'
gdata: to automatically download and install the perl
gdata: libraries needed to support Excel XLS and XLSX formats.

Attaching package: 'gdata'

The following objects are masked from 'package:dplyr':

combine, first, last

The following object is masked from 'package:purrr':

```
keep
```

The following object is masked from 'package:stats':

```
nobs
```

The following object is masked from 'package:utils':

```
object.size
```

The following object is masked from 'package:base':

```
startsWith
```

```
library(httr)
```

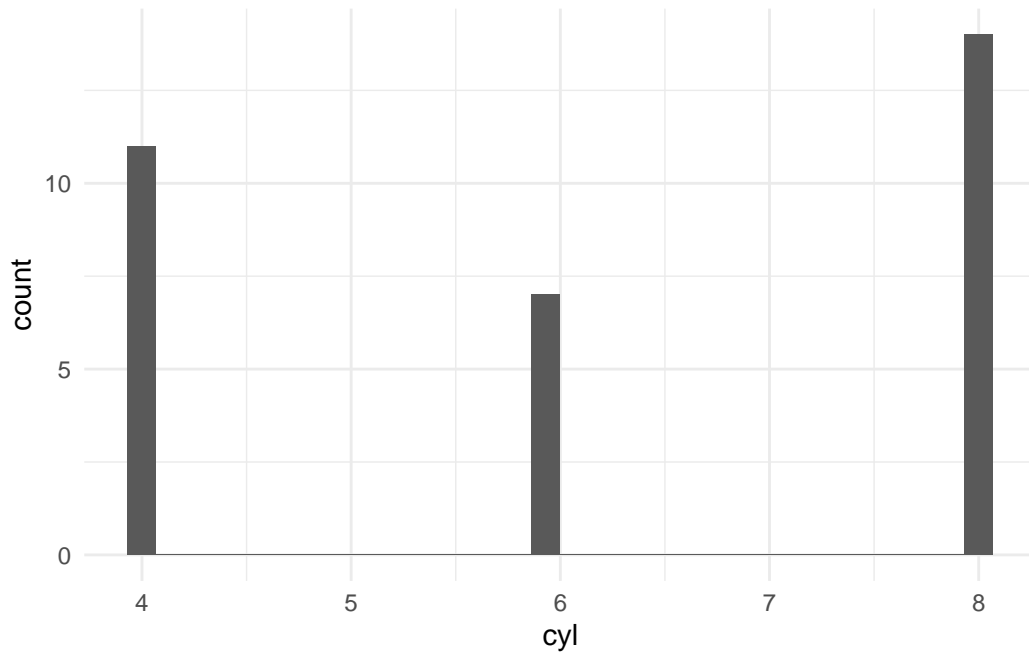
Oppgave I: Introduksjon til histogram

Et histogram eller frekvensfordeling er en figur som viser hvor ofte forskjellige verdier oppstår i et datasett. Frekvensfordelinger spiller en grunnleggende rolle i statistisk teori og modeller. Det er avgjørende å forstå de godt. En kort innføring følger.

La oss se på et eksempel. I datasettet `mtcars` viser variabelen `cyl` antall sylindere i motorene til kjøretøyene i utvalget.

```
data(mtcars)
mtcars %>%
  ggplot(aes(cyl)) +
  geom_histogram() +
  theme_minimal()
```

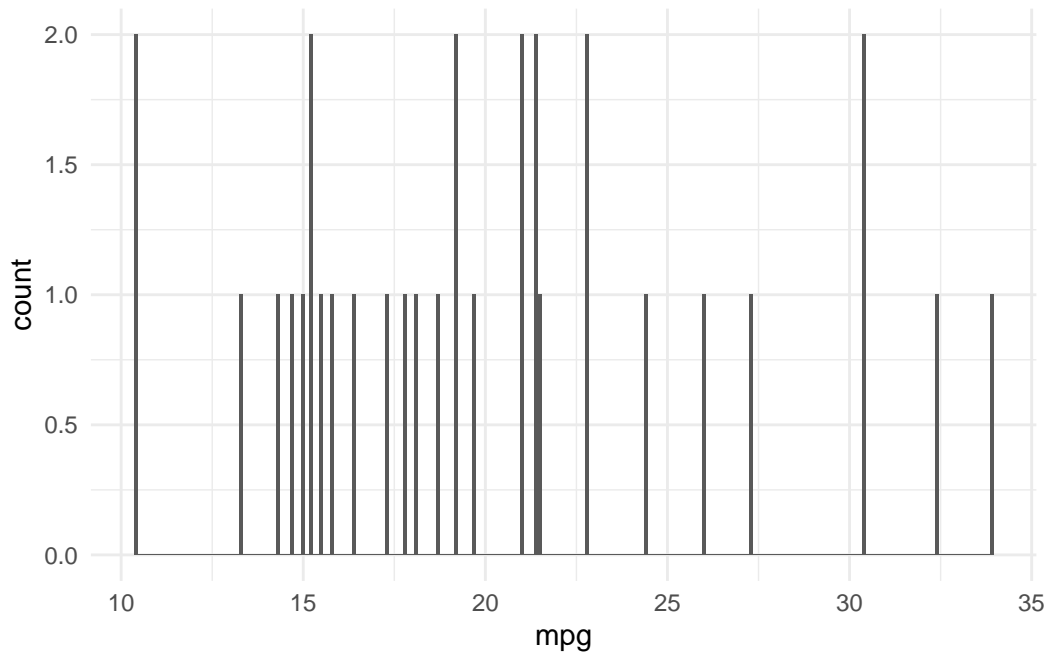
``stat_bin()`` using ``bins = 30``. Pick better value with ``binwidth``.



Verdiene av variabelen er gitt ved den horisontale aksen, antall observasjoner på den vertikale aksen. Vi ser at det er 11, 7, og 14 biler med henholdsvis 4, 6, og 8 sylindere.

La oss betrakte et eksempel til. Variabelen `mpg` i `mtcars` måler gjennomsnittlig drivstofforbruk i uanstendige engelske enheter. Variabelen er målt med ett desimal i presisjon.

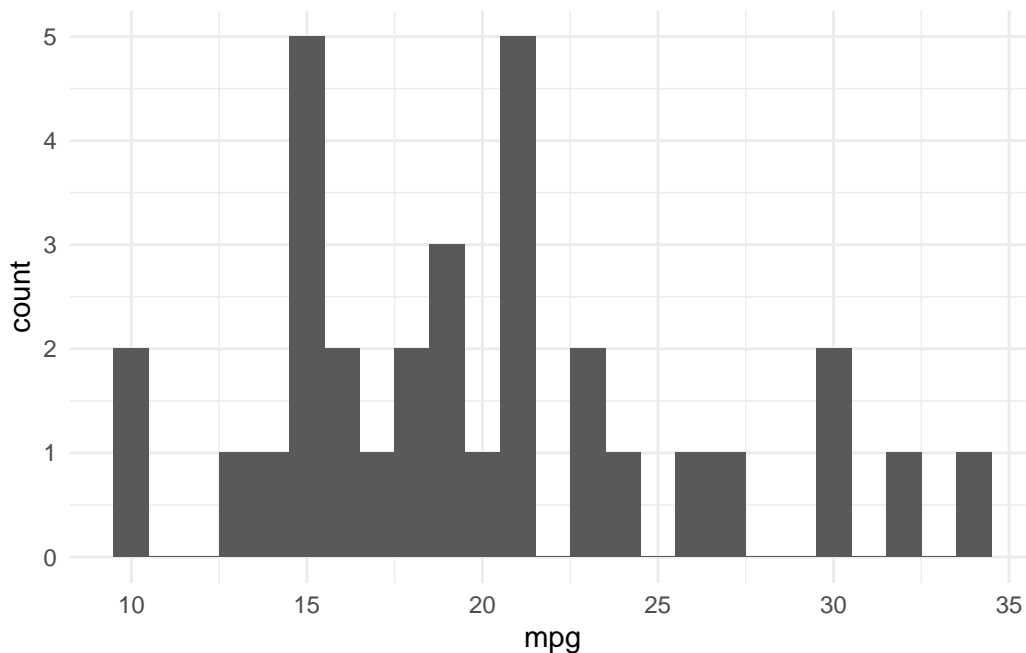
```
data(mtcars)
mtcars %>%
  ggplot(aes(mpg)) +
  geom_histogram(binwidth=0.1) +
  theme_minimal()
```



Datasettet inneholder mange unike verdier, hvilket gir utslag i et flatt histogram, noe som er lite informativt. Løsningen da er å gruppere verdier som ligger i nærheten av hverandre. Kommandoen `binwidth` i `geom_histogram()` bestemmer bredden av intervallene som blir slått sammen. Kan du forklare hvorfor alle unike verdier blir telt ved å bruke `binwidth = 0.1`?

Ekspirer med forskjellige verdier for `binwidth` og forklar hva som kjennetegner en god verdi.

```
# løs oppgave I her
data(mtcars)
mtcars %>%
  ggplot(aes(mpg)) +
  geom_histogram(binwidth=1) +
  theme_minimal()
```



#En god verdi er en verdi som gjør det enkelt å lese og tolke dataene samtidig som du ikke

Oppgave II: Last ned og rydd i data

Vi skal nå undersøke dataene i [Tabell 07967: Kostnader til egenutført FoU-aktivitet i næringslivet, etter næring \(SN2007\) og sysselsettingsgruppe \(mill. kr\) 2007 - 2020 SSB](#). Dere skal laste de ned ved hjelp av API. Se [brukerveiledningen](#) her.

Bruk en JSON-spørring til å laste ned alle statistikkvariable for alle år, næringer, og sysselsettingsgrupper med 10-19, 20-49, 50-99, 100-199, 200 - 499, og 500 eller flere ansatte. Lagre FoU-kostnader i milliarder kroner. Sørg for at alle variabler har riktig format, og gi de gjerne enklere navn og verdier der det passer.

Hint. Bruk lenken til SSB for å hente riktig JSON-spørring og tilpass koden fra case 3.

```
# besvar oppgave II her
url <- "https://data.ssb.no/api/v0/no/table/07967/"

query <- '{
  "query": [
    {
```

```

"code": "NACE2007",
"selection": {
  "filter": "item",
  "values": [
    "A-N",
    "C",
    "G-N",
    "A-B_D-F"
  ]
},
{
  "code": "SyssGrp",
  "selection": {
    "filter": "item",
    "values": [
      "10-19",
      "20-49",
      "10-49",
      "50-99",
      "100-199",
      "200-499",
      "500+"
    ]
  }
},
{
  "code": "Tid",
  "selection": {
    "filter": "item",
    "values": [
      "2007",
      "2008",
      "2009",
      "2010",
      "2011",
      "2012",
      "2013",
      "2014",
      "2015",
      "2016",

```

```

        "2017",
        "2018",
        "2019",
        "2020"
      ]
    }
  }
],
"response": {
  "format": "json-stat2"
}
}'

hent_indeks.tmp <- url %>%
  POST(body = query, encode = "json")

df <- hent_indeks.tmp %>%
  content("text") %>%
  fromJSONstat() %>%
  as_tibble()

df$år <- as.integer(df$år)
df <- rename(df, verdi = value)
df <- rename(df, næring = `næring (SN2007)`)

```

Oppgave III: Undersøk fordelingen

Vi begrenser analysen til bedrifter med minst 20 ansatte og tall fra 2015 - 2020. Lag en figur som illustrerer fordelingen av totale FoU-kostnader fordelt på type næring (industri, tjenesteyting, andre) og antall ansatte i bedriften (20-49, 50-99, 100-199, 200-499, 500 og over). Tidsdimensjonen er ikke vesentlig, så bruk gjerne histogram.

Merknad. Utfordringen med denne oppgaven er at fordelingene er betinget på verdien av to variable. Kommandoen `facet_grid()` kan være nyttig til å slå sammen flere figurer på en ryddig måte.

```

# besvar oppgave III her
df_ny <- df %>%
  filter(sysselsettingsgruppe!="10-19 sysselsatte") %>%
  filter(sysselsettingsgruppe!="10-49 sysselsatte") %>%

```



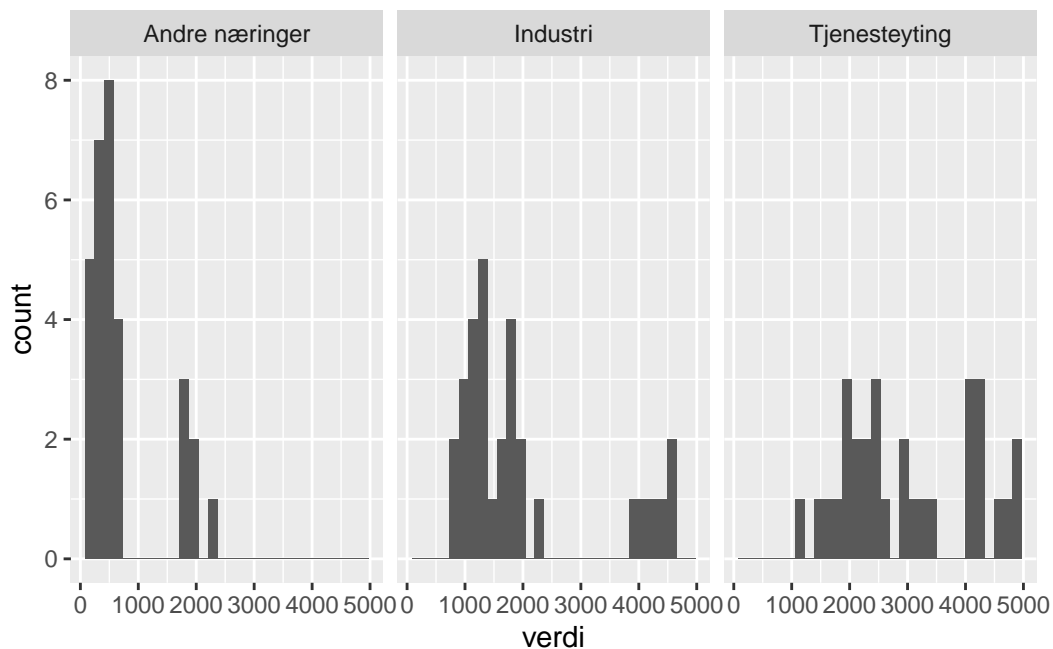
```

filter(statistikkvariabel == "FoU-kostnader i alt") %>%
filter(næring!="Alle næringer") %>%
filter(år %in% c(2015:2020))

df_ny %>%
  ggplot(aes(verdi)) +
    geom_histogram() +
    facet_grid(~næring)

```

`stat_bin()` using `bins = 30`. Pick better value with `binwidth`.



Oppgave IV: Undersøk fordelingen igjen

Kan du modifisere koden fra oppgave II til å i tillegg illustrere fordelingen av FoU-bruken på lønn, innleie av personale, investering, og andre kostnader?

Merknad. Kommandoen `fill = [statistikkvariabel]` kan brukes i et histogram.

```

# besvar oppgave III her
df_ny <- df %>%

```

```

filter(sysselsettingsgruppe!="10-19 sysselsatte") %>%
filter(sysselsettingsgruppe!="10-49 sysselsatte") %>%
filter(næring!="Alle næringer") %>%
filter(år %in% c(2015:2020))

```

```

df_ny %>%
  ggplot(aes(verdi)) +
  geom_histogram(aes(fill = statistikkvariabel),binwidth=500) +
  facet_grid(~næring)

```

