



PSY2014 – KVANTITATIV METODE

Forelesning 1: Grunnleggende konsepter, og dataanalyse i R
Nikolai Czajkowski

OPPTAK

- **Det gjøres opptak av forelesningen**
- Opptaket vil bli lagret på emnesiden til PSY2014 UiO, og en lenke i canvas vil tilgjengelig for de som følger kurset.
 - <https://uiuo.instructure.com/>
- Opptaket skal bli slettet etter 2023.

- **Nikolai Czajkowski** (kontor S04-06 n.o.czajkowski@psykologi.uio.no).
- Gjennom kurset tar vi opp flere av de viktigste grunnleggende statistiske analysene.
- I all hovedsak er det et *anvendt* kurs, men jeg vil gi en viss innføring i teorien også.
- Problemløsing er essensielt i PSY2014.
 - Det viktigste er ikke å huske definisjoner / formler.
 - Ikke begynn en uke før eksamen, det tar litt tid å synke inn.
- Spør, veldig gjerne på forelesning, men i alle fall på seminar!

SEMINARØVELSER

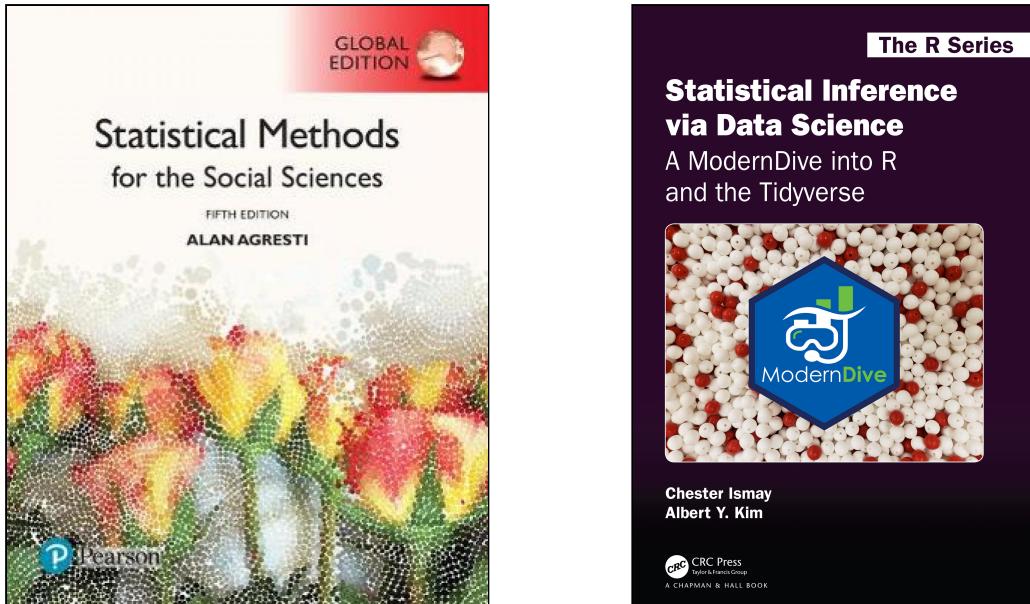
- 8 uker med seminarer (begynner neste uke).
- Seminargruppene blir ikke podcastet.
- Det forutsettes at dere har sett på oppgavene før seminaret.
 - Løsningsforslag vil bli lagt ut i slutten av uken.
- Oppgavene har to deler:
 1. Teoretisk del som vil bli løst med hoderegning/kalkulator.
 2. Praktisk data-analyse i R.

PORTALEN TIL OPPGAVESETTENE

- Ukeoppgaver blir tilgjengelig på følgende nettside:
 - <https://shinyibv02.uio.no/content/bedbf8ff-2a80-4521-9ddb-5bcd6e7ca6b8>
- Her finner dere både:
 - Teoretiske oppgaver som gjennomgås på seminarøvelsene.
 - Interaktive nettbaserte oppgaver som gjor en innføring til data-analyse i R.

EKSAMEN

- 23 . mai kl. 09:00 (3 timer), Silurveien 2 Sal 3D.
- Oppgavene vil i likne på øvelsene; Enkle utregninger, tolkning av R-utskrift og diskusjon opp mot pensum.
- Alle relevante formler blir vedlagt oppgaven.
- Det vil bli mulig å benytte en enkel kalkulator under eksamen.
- Se eksamensoppgaver i PSY2014 (og PSY2011 og PSY2012) på nett:
 - <http://www.sv.uio.no/studier/ressurser/tidligere-eksamensoppgaver/psykologi.html>
- Løsningsforslag noen tidligere eksamensoppgaver vil bli lagt ut i andre halvdel av semesteret.



- Agresti, Alan. Statistical Methods for the Social Sciences: Pearson New International Edition (4th or 5th edition) ISBN-13: 978-0134507101
 - Kapitler: (1-6) 7 - 14

Støttelitteratur for data-analyse i R:

- Ismay & Kim, A Modern Dive into R and the Tidyverse (2022).
 - <https://moderndive.com/>

FORELESNINGSPLAN

Nr	Dato	tid	Foreleser	Tema	Agresti
1	08. feb.	14:15–16:00	NOC	Grunnleggende kvantitative konsepter	Kap 1-4
2	15 feb.	14:15–16:00	NOC	Bivariat regresjon	Kap 9
3	01. mars.	14:15–16:00	NOC	Multipel regresjon	Kap (10), 11
4	08. mars.	14:15–16:00	NOC	Inferens i regresjon	Kap (5, 6), 9.5, 10.5
5	15. mars	14:15–16:00	NOC	Kategoriske prediktorer, interaksjon, mediering	11.4, 13, 14
6	22. mars	14:15–16:00	NOC	Kji-kvadrat analyser	Kap 8.1, 8.2, 8.3
7	29. mars	14:15–16:00	DEE	Variansanalyse	(Kap 7,12,13)
8	05 april	14:15–16:00	DEE	Variansanalyse	(Kap 7,12,13)
9	19. april	14:15–16:00	NOC	Eksamensoppgaver/Repetisjon	

Jeg vil sannsynligvis sette opp noen ekstra Q&A zoom timer, disse blir annonseret etter hvert.

VARIABLER

- Kvantitative metoder er basert på antagelsen om at egenskaper kan kodes i variabler.
- For å være interessante må variabler *variere*, og hensikten med kvantitative metoder er å redegjøre for hva variasjonen skyldes.
- Vi skiller ofte mellom
 - **Avhengig** variabel / respons.
 - **Uavhengig** variabel / prediktor / forklaringsvariabel.
- Forskjellige variabler kan ta ulike typer verdier:
 - Kontinuerlige (høyde inntekt),
 - Kategorisk/Faktor (ordinal; vektklasser, nominal; bilmerker).
- Et **datasett** er en samling av en eller flere observerte skårer på en eller flere variabler.

UTVALG OG POPULASJONER

- **Populasjon:** De du ønsker å ønsker å uttale deg om.
 - Ikke tenk på populasjonen som en faktisk gruppe, vi vil representere den som en fordeling vi trekker observasjoner fra.
- **Utvalg (sample):** Et relativt lite antall observasjoner trukket fra populasjonen.
- **Statistikk:** En egenskap (verdi) regnet ut på bakgrunn av et utvalg.
 - **Beskrivende statistikk:** (Descriptive statistics) Kvantifiserer egenskaper ved utvalget.
 - **Slutningsstatistikk** (Inferential statistics): Metoder som lar oss å trekke slutsninger om populasjonen gjennom å se på data fra et utvalg.
 - Krever en statistisk modell, dvs. et sett av *antagelser* om hvordan observasjonene ble generert.

OBSERVASJONELLE VS. EKSPERIMENTELLE STUDIER

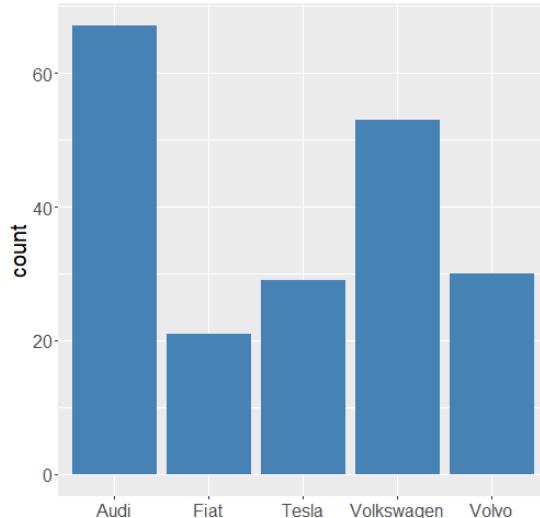
- **Eksperimentelle studier**

- Deltagere (randomisert) inn i eksperiment eller kontrollgruppe
- Forskeren gjennomfører en behandling/intervensjon, og observerer så effekten.
- *Tradisjonelt analysert med variansanalyse*

- **Observasjonelle studier**

- Deltagere får selv selektere seg inn i grupper, og forskeren påfører ingen intervensjon.
- *Tradisjonelt analysert med regresjon*

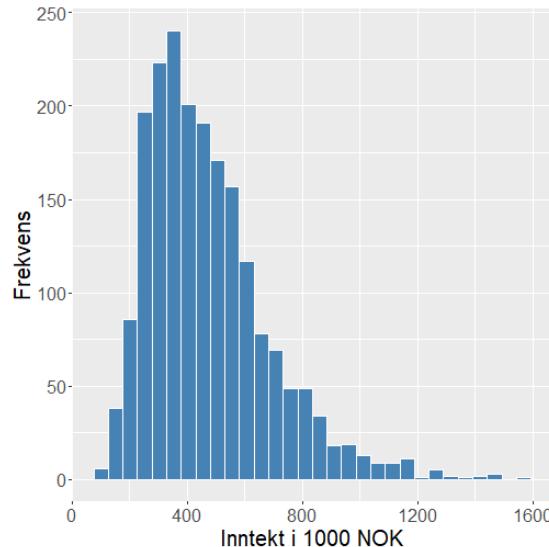
GRAFER AV KATEGORISKE VARIABLER



- **Frekvensfordeling** (frequency distribution): En liste mulige verdier en variabel kan ta, og antall observasjoner som tar hver verdi.
- Kategoriske varabler blir typisk avbildet med et stolpediagram (bar chart).

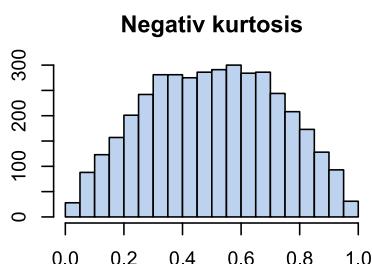
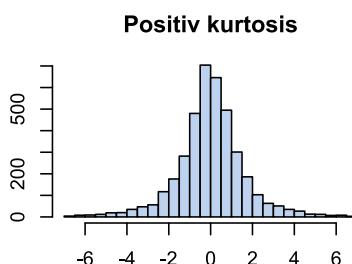
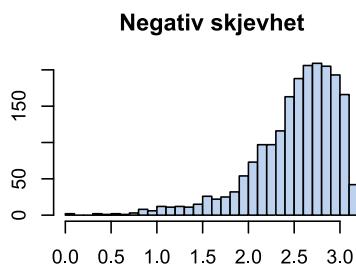
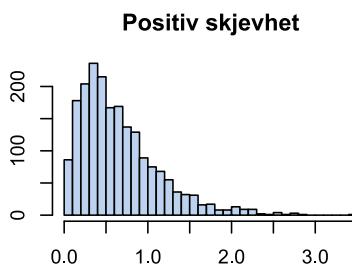
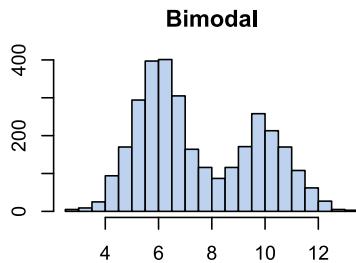
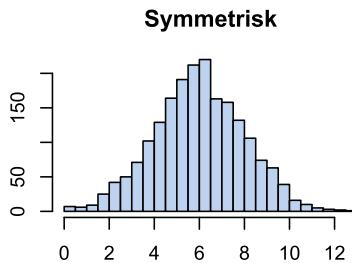
GRAFER AV KONTITUERLIG DATA (A: 3.1)

Histogram: del rangen inn i intervaller, tell antall observasjoner som faller i hvert intervall, og lag et stolpediagram.



- Hva er den generelle formen på fordelingen?
- Hva er typiske verdier?
- Er det mye variabilitet?
- Er det atypiske verdier?

HVORDAN BESKRIVE FORDELINGER



Modalitet: Antall topper i histogrammet

Skjevhet (skew): Mål på grad av asymmetri.

Kurtosis: Den relative koncentrasjonen av skårer i sentrum versus halene av fordelingen.

SIGMA NOTASJON

- Anta at du har et *ordnet* (dvs. rekkefølgen er av betydning) sett av verdier:

$$X = \{2, 5, -1, 11\}$$

- Vi referer til enkeltelementer i settet med et *subskript*:
 - $x_2 = 5$
- For å vise til *summen* av et sett av elementer bruker vi det greske "Sigma" (sum) tegnet:
 - $\sum_{i=1}^n x_i$: Spesifisering av øvre og nedre grense er mest presist, men krever mer knot.
 - Σx Når øvre og nedre grense ikke er spesifisert menes "summen av alle elementene i settet X".
- Eksempler:
 - $\sum_{i=1}^n x_i = 2 + 5 + (-1) + 11 = 17$
 - $\sum_{i=1}^n (x_i)^2 = 2^2 + 5^2 + (-1)^2 + 11^2 = 151$
 - $\sum_{i=1}^n (x_i - 2)^2 = (2 - 2)^2 + (5 - 2)^2 + (-1 - 2)^2 + (11 - 2)^2$

MÅL PÅ SENTRALITET (A:3.2)

Gitt settet $X = \{1, 2, 3, 4, 5, 8, 8, 20, 100\}$

Mode: Verdien som opptrer flest antall ganger.

- Mode av X er 8.

Median: Verdien som splitter en fordeling/ordnet sett inn i to like store deler.

- Medianen til X er 5.
- Dersom settet har partall elementer er median definert som gjennomsnittet av de to midterste tallene.

Gjennomsnittet (mean):

- $\bar{X} = \frac{\sum X}{n}$
- $\bar{X} = \frac{1}{9}(1 + 2 + 3 + 4 + 5 + 8 + 8 + 20 + 100) = 16.77$

GJENNOMSNITT SOM BALANSEPUNKT

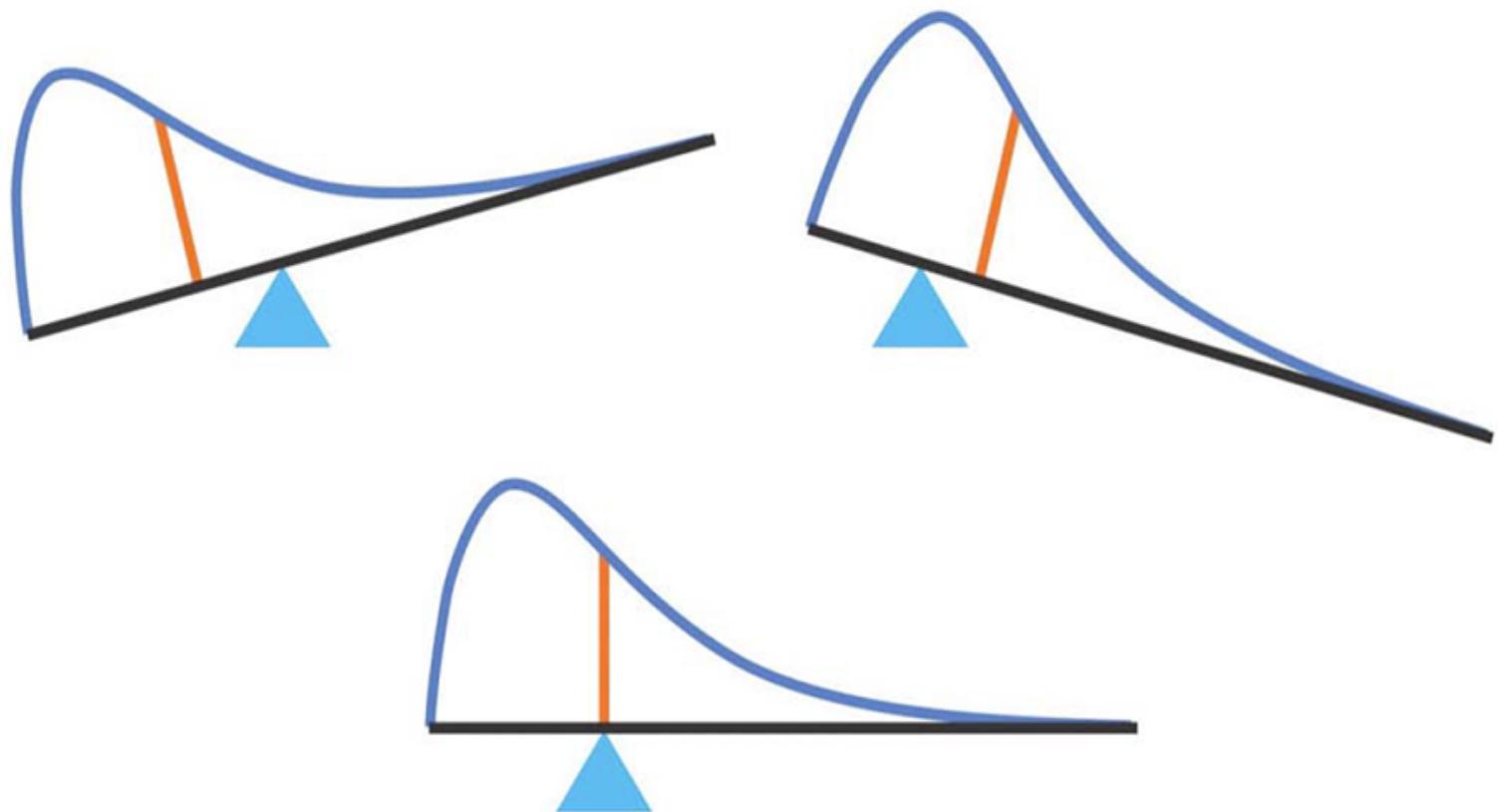
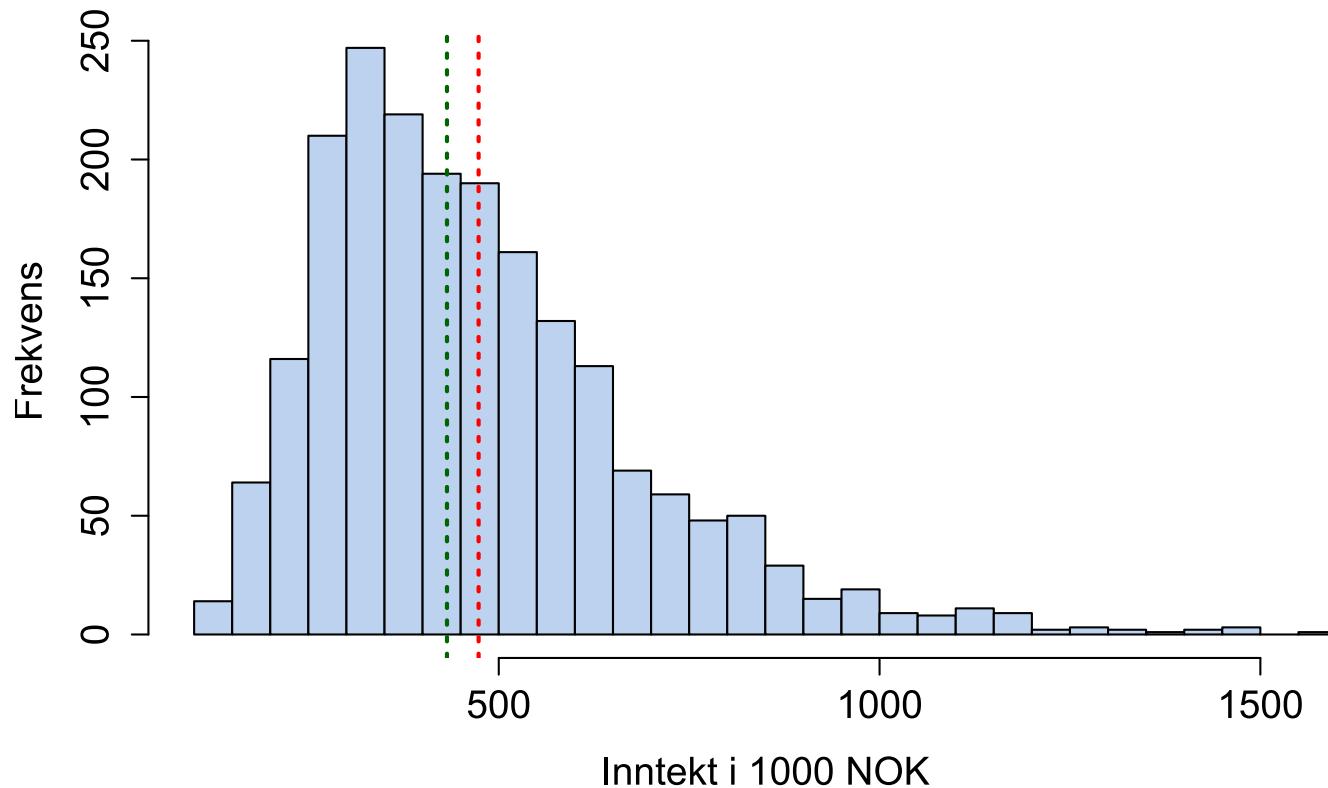


Figure 1-25

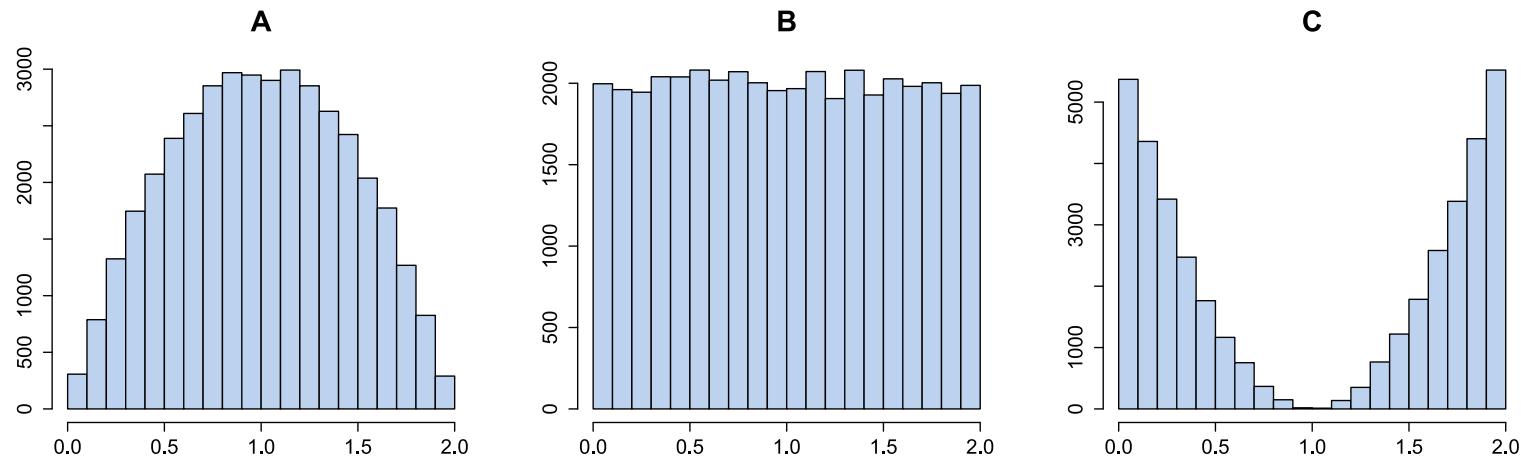
Introduction to the Practice of Statistics, Fifth Edition

© 2005 W.H. Freeman and Company

HVILKEN LINJE ER MODE, MEDIAN OG GJENNOMSNITTET AV FORDELINGEN?



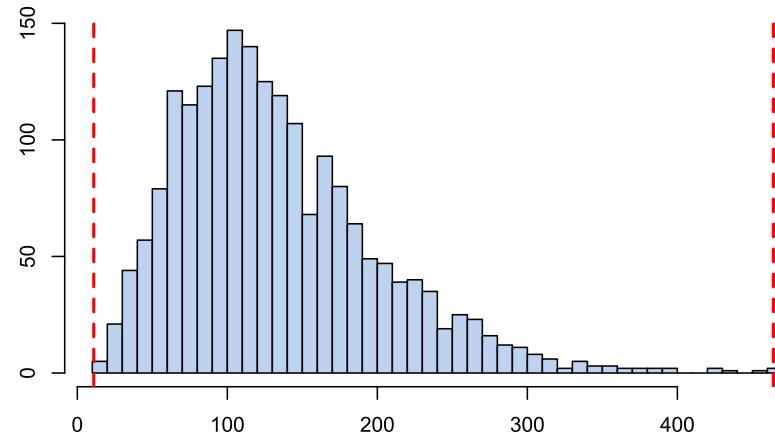
VARIANS



$$Var(X) = s_X^2 = \frac{\sum(X_i - \bar{X})^2}{n - 1}$$

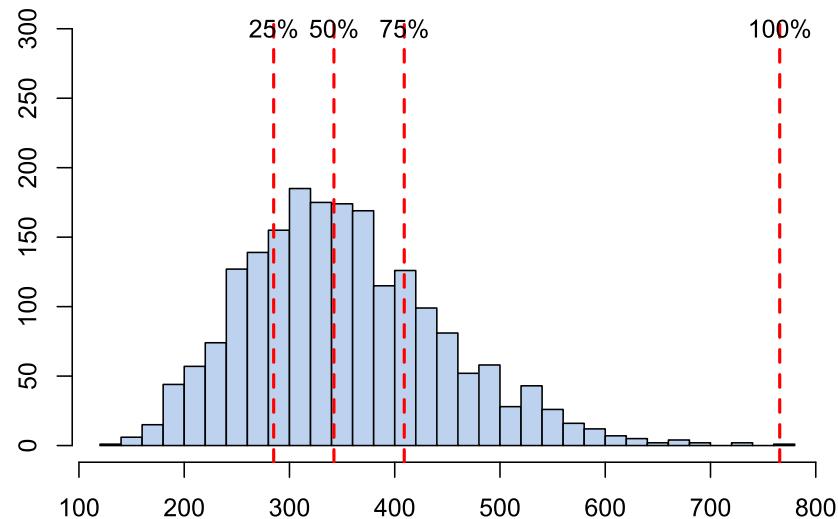
Varians er et mål på variabilitet rundt den forventede verdien (gjennomsnittet) på en fordeling.

MÅL PÅ VARIABILITET (A:3.3)



- **Range:** maksimum verdi – minimum verdi

PROSENTILER



Posisjonsmål forteller oss hvor på fordelingen en observasjon ligger.

- **Prosentil (percentile):** Den p'te prosentil er den verdien som en gitt prosent er mindre enn eller lik.
- **Kvartil (quartile):** Verdier som deler det sorterte datasettet inn i fire like store grupper, Q1,Q2,Q3 and Q4.
- **Interkvartil range (IQR):** 75. prosentil – 25 . prosentil (Q3-Q1)

STANDARDAVVIK

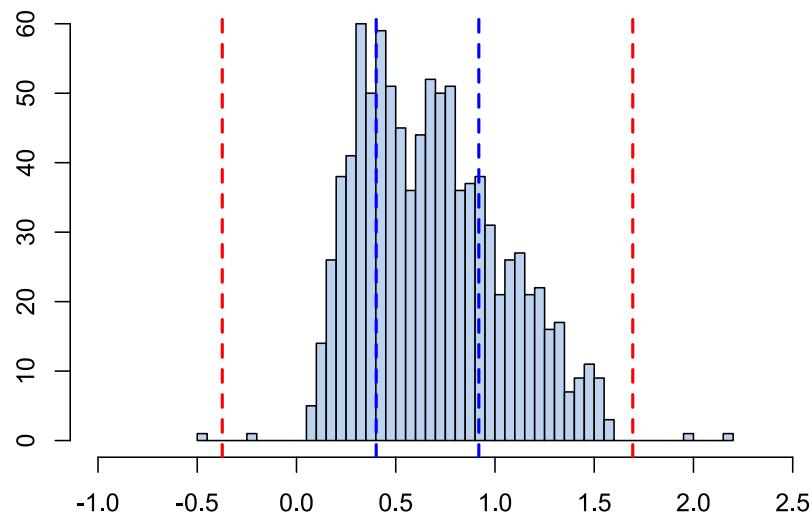
$$Var(X) = s_X^2 = \frac{\sum(X_i - \bar{X})^2}{n - 1}$$

$$Sd(X) = s_X = \sqrt{s_X^2}$$

Egenskaper til standardavviket

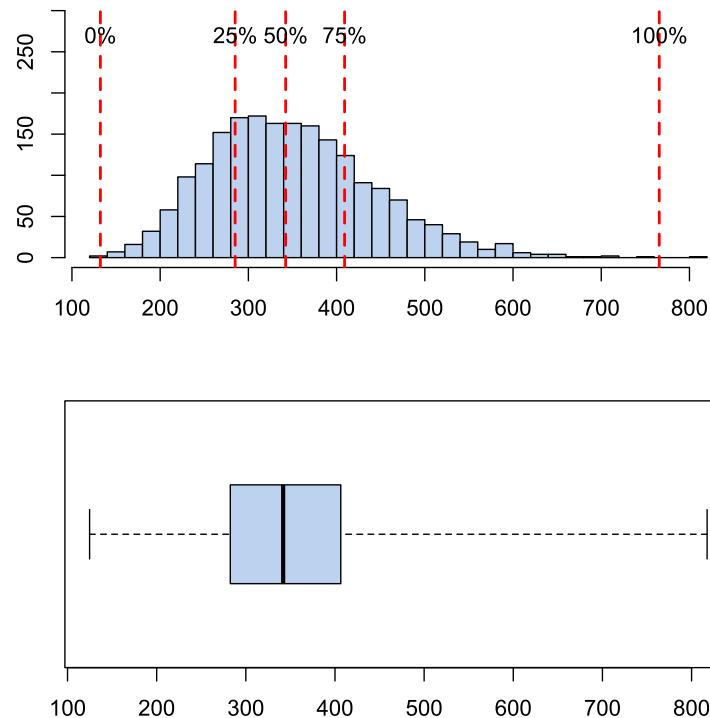
- Et spredningsmål med samme enheter som variabelen
 - $s_X \geq 0$, og bare 0 dersom alle verdiene er like.
- Kan tolkes som "gjennomsnittlig avvik fra gjennomsnittet"
 - Vi deler på $n-1$ fordi snittet må estimeres samtidig (mer om det senere).
- Varians of standardavvik er lite robust, (dvs. sterkt påvirket av uteliggere).

UTELIGGERE



- **Uteligger (outlier):** En observasjon som ligger langt i verdi fra de øvrige observasjonene
 - "Langt" er typisk definert som verdier større eller lik $Q3 + (1.5 \cdot IQR)$, eller mindre enn $Q1 - (1.5 \cdot IQR)$

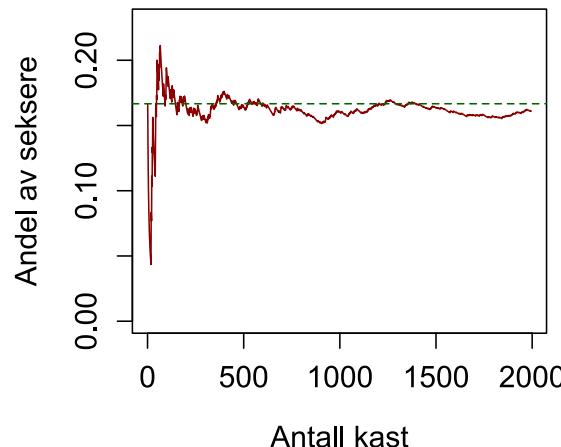
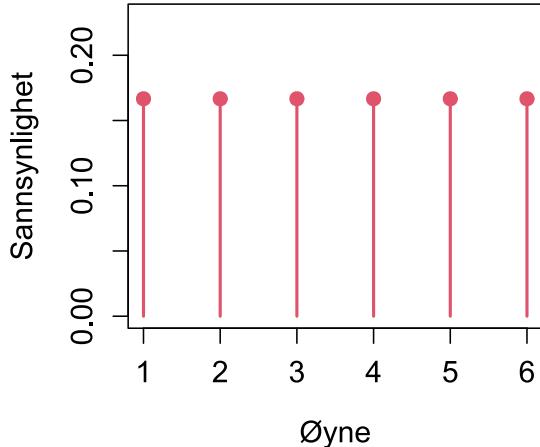
5 –TALLS OPPSUMMERING AV EN FORDELING



Boxplot er ofte en god måte å vise viktige egenskaper ved fordelingen.

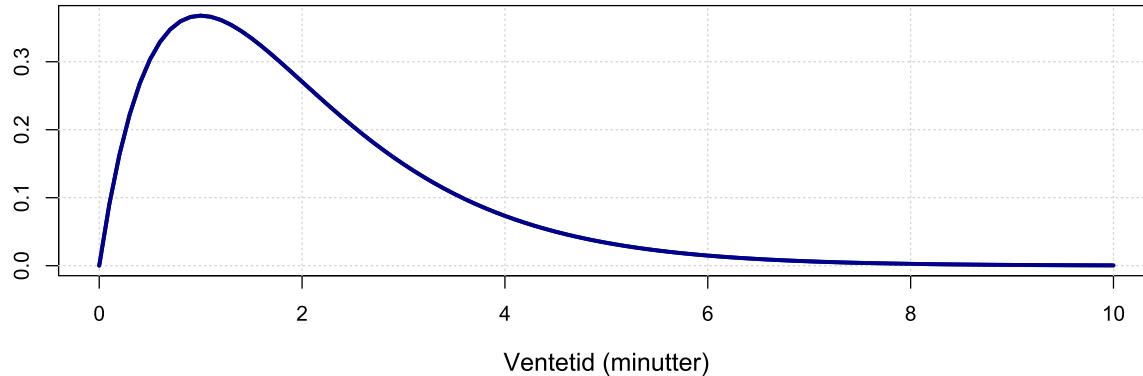
SANNSYNLIGHET

Utfall av terning



- Et **stokastisk utfall** er tilfeldig, men følger en regularitet over mange repetisjoner.
- **Sannsynlighet** kan definers som andelen ganger et gitt utfall forekommer over et stort antall repetisjoner.
 - Følgelig tar sannsynlighet en verdi mellom 0 og 1.

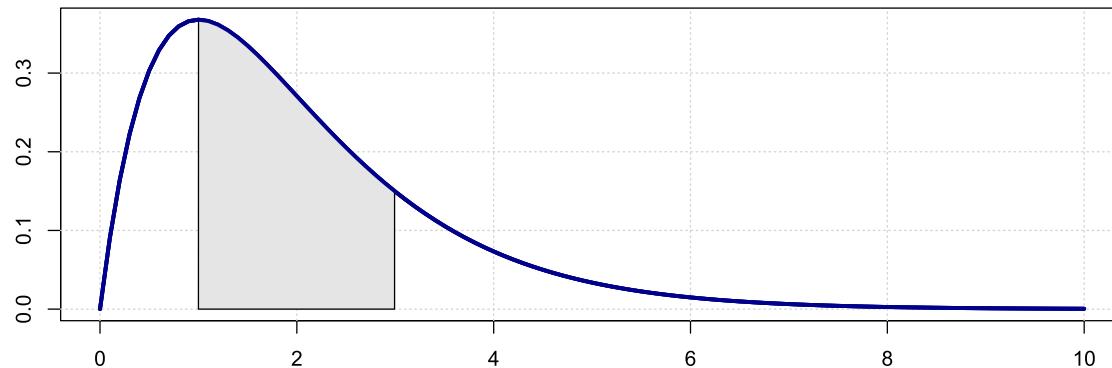
SANNSYNLIGHETSFORDELINGER FOR KONTINUERLIGE UTFALL



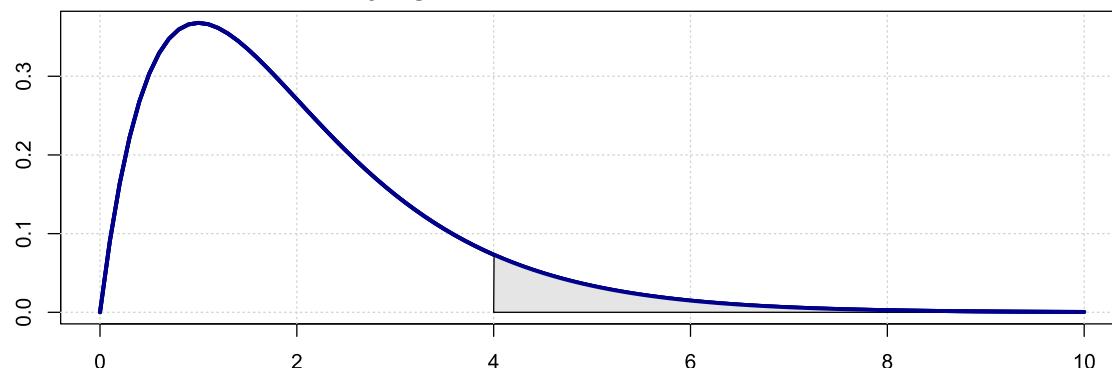
- **En sannsynlighetstetthet** (probability density) er en funksjon som viser den relative forekomsten av utfall av en stokastisk variabel som tar kontinuerlige verdier.
 - Det totale arealet under kurven er lik 1.
 - Ingen verdier er negative
 - Sannsynligheten at den stokastiske variabelen tar verdier i et gitt intervall er lik arealet under kurven for dette intervallet.

HVORDAN FINNE SANNSYNLIGHET FOR ULIKE UTFALL?

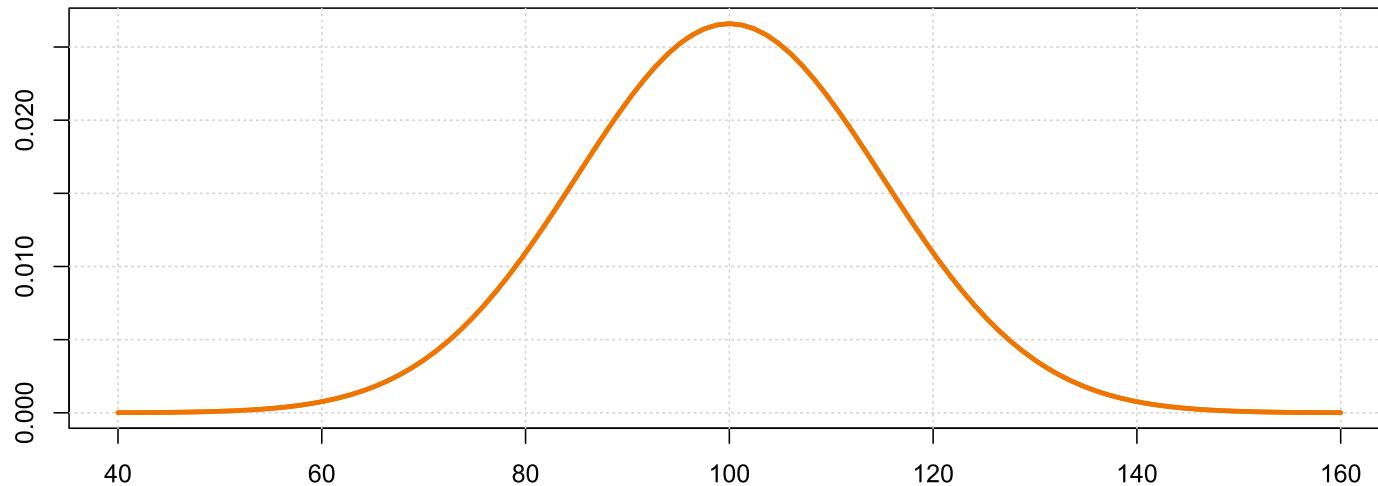
Hva er sannsynligheten for å vente mellom 1 og 3 minutter? ca. 0.54



Hva er sannsynligheten for å vente mer enn 4 minutter? ca. 0.09

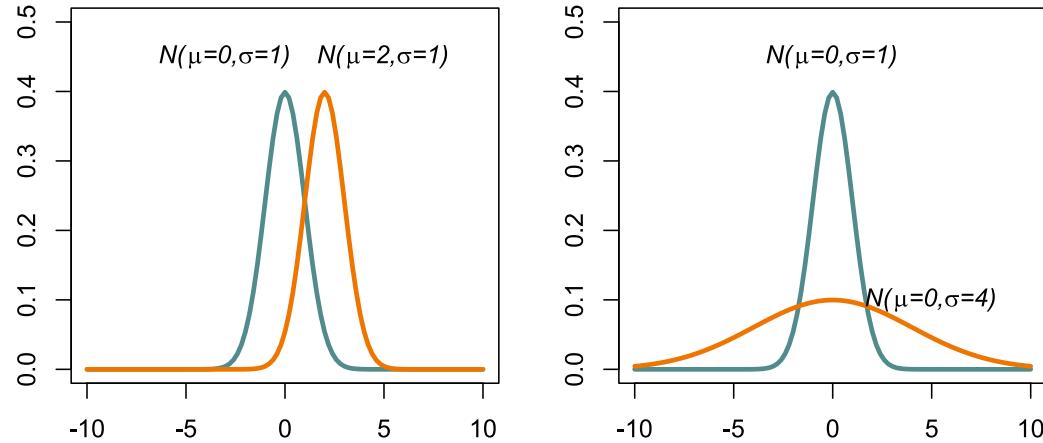


NORMALFORDELINGEN



Normalfordelingen er kanskje den viktigste fordelingen innen statistikk.

NORMALFORDELINGEN (2)

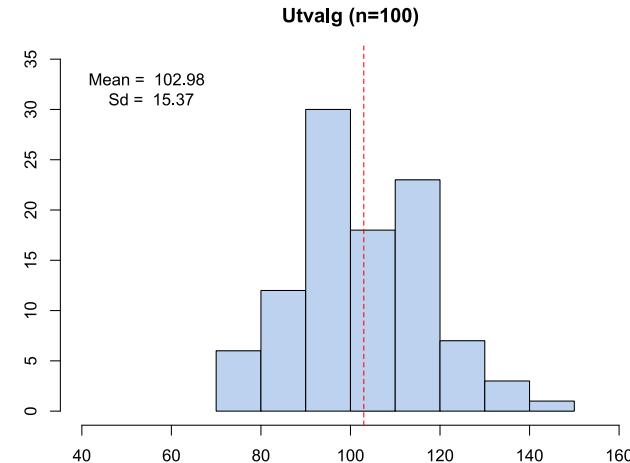
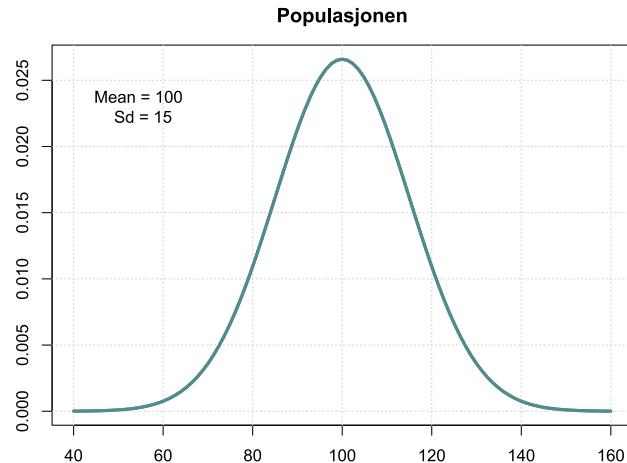


At en stokastisk variabel X er normalfordelt skriver vi ofte forkortet $X \sim N(\mu, \sigma^2)$.

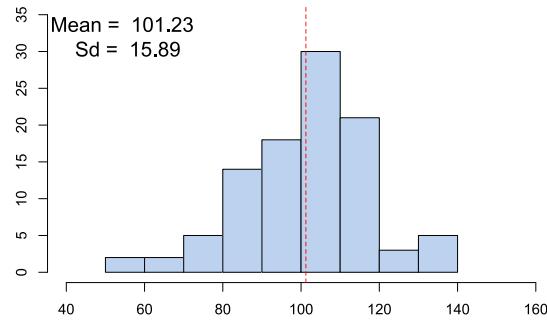
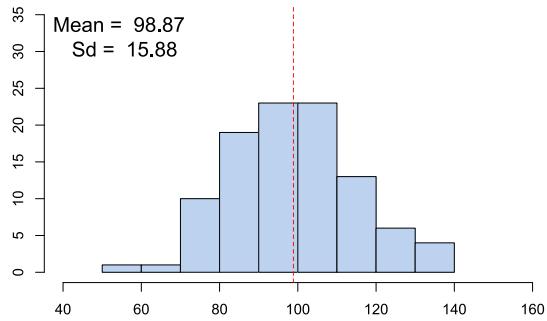
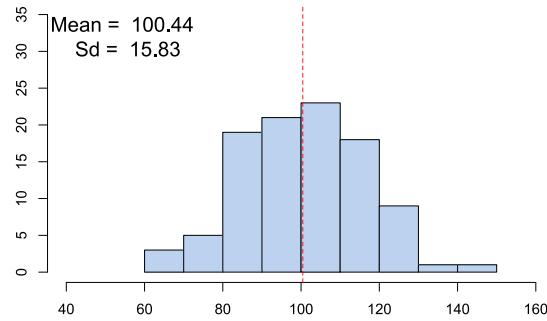
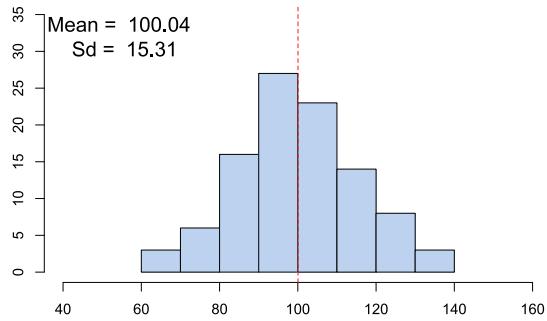
Uttrykket for normalfordelingen er gitt ved:

$$f(x|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(x-\mu)^2}{2\sigma^2}}$$

STATISTIKKER VS PARAMETERE

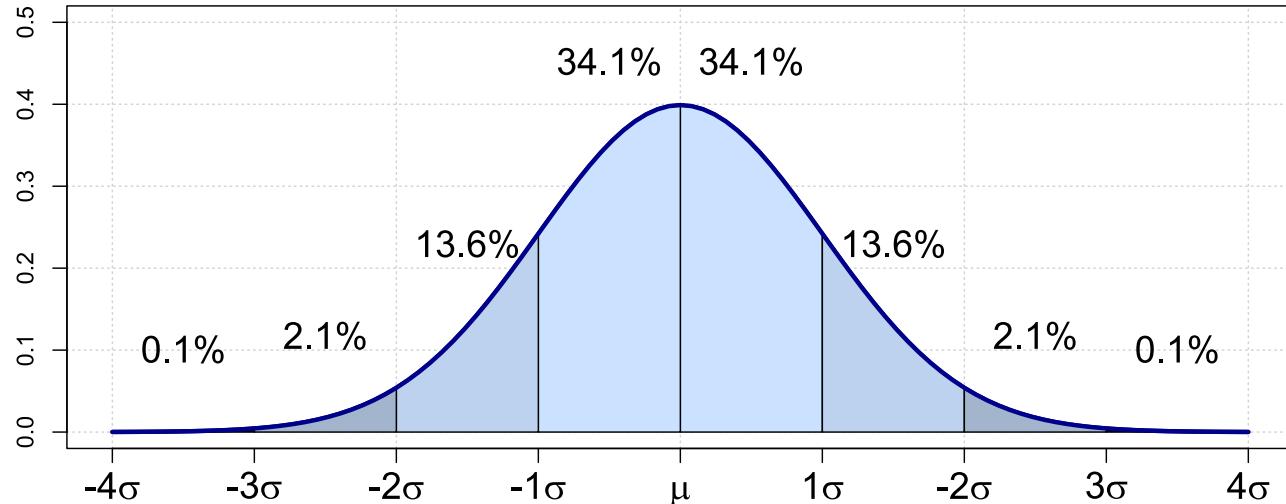


- Vi antar ofte at egenskapen vi er interessert i følger en bestemt fordeling i populasjonen, og fra denne fordelingen trekker vi et utvalg.
- Verdiene som bestemmer formen på fordelingen i populasjonen kalles **parametere**.
 - Feks. IQ er normalfordelt i populasjonen med snitt 100, of sd 15. 100 og 15 er her **parametere**.
- Verdier beregnet i et konkret utvalg kalles **statistikker**.
 - I utvalget til høyre (n=100), trukket fra populasjonen er snittet 99.31 og standardavviket 16.1.



Deskriptive statistikker vil ta ulike verdier i forskjellige utvalg trukket fra den samme populasjonen.

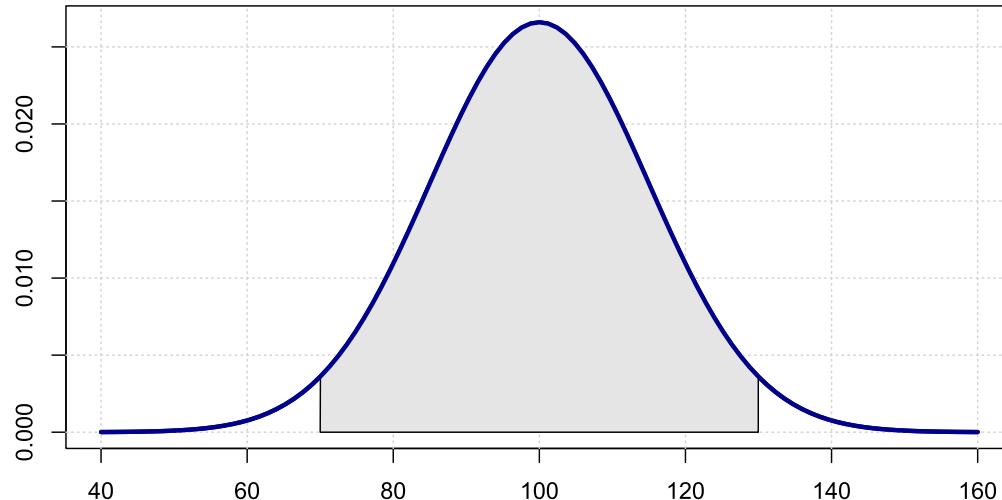
AREALER AV NORMALFORDELINGEN



Alle normalfordelinger, uansett form har følgende egenskap:

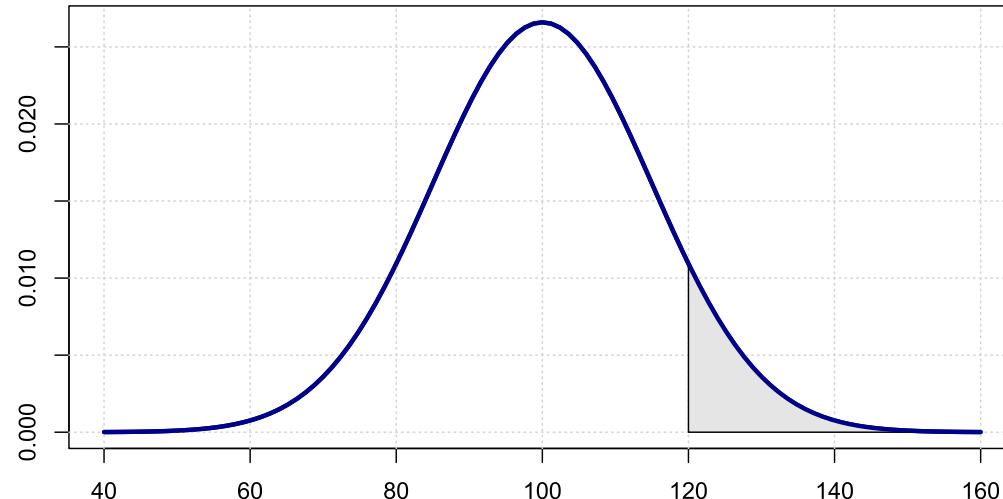
- Ca. 68% av fordelingen ligger innenfor ett standardavvik fra snittet.
- Ca. 95% av fordelingen ligger innenfor to standardavvik fra snittet.
- Ca. 99.7% av fordelingen ligger innenfor ett standardavvik fra snittet.

SANNSYNLIGHET AV INTERVALLER



Hva er sannsynligheten for at en tilfeldig valgt person har IQ mellom 70 og 130?

STANDARD NORMALFORDELING



- Hva om du trenger sannsynligheten av et intervall som ikke er et helt antall standardavvik?
 - For eksempel. Hva er sannsynligheten for at en person har IQ større enn 120?
- For å finne sannsynlighet av intervaller regner vi som regel om til **Z-skåre**.

Z-SKÅRE

Z-skåre: Omregning av variabelen til enheter som utgjør antall standardavvik fra gjennomsnittet.

$$Z = \frac{X - \bar{X}}{s_X}$$

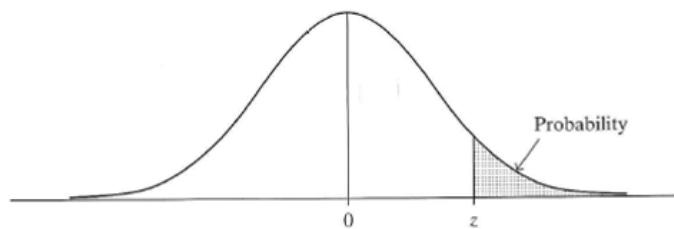
For eksemplet på forrige slide:

$$Z = \frac{X - 100}{15} = \frac{120 - 100}{15} = 1.333$$

- Viktig: Dersom en variabel $X \sim N(\mu, \sigma^2)$, er $Z \sim N(0, 1)$
 - dvs. en normalfordeling med snitt 0 og standardavvik 1.

BRUK AV TABELL (AGRESTI TABLE A)

TABLE A: Normal curve tail probabilities. Standard normal probability in right-hand tail (for negative values of z , probabilities are found by symmetry).



z	Second Decimal Place of z									
	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.1	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.2	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.3	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.4	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.5	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.6	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.7	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.8	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.9	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.0	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.1	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.2	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.3	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0722	.0708	.0694	.0681
1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.8	.0359	.0352	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233

THE R PROJECT FOR STATISTICAL COMPUTING

- R er et programmeringsspråk og et gratis software miljø for statistiske beregninger og grafikk, støttet av "the R Foundation".
 - <https://www.r-project.org/>
- Språket R er *veldig* populært blant statistikere, data-scientists og data-miners for utvikling av statistisk programvare og analyse.
 - Kalles R delvis som et vink til de første to forfatterne, og delvis til et språk "S".
 - Kommer med et lite sett av funksjoner, men kan utvides fra et enormt bibliotek.

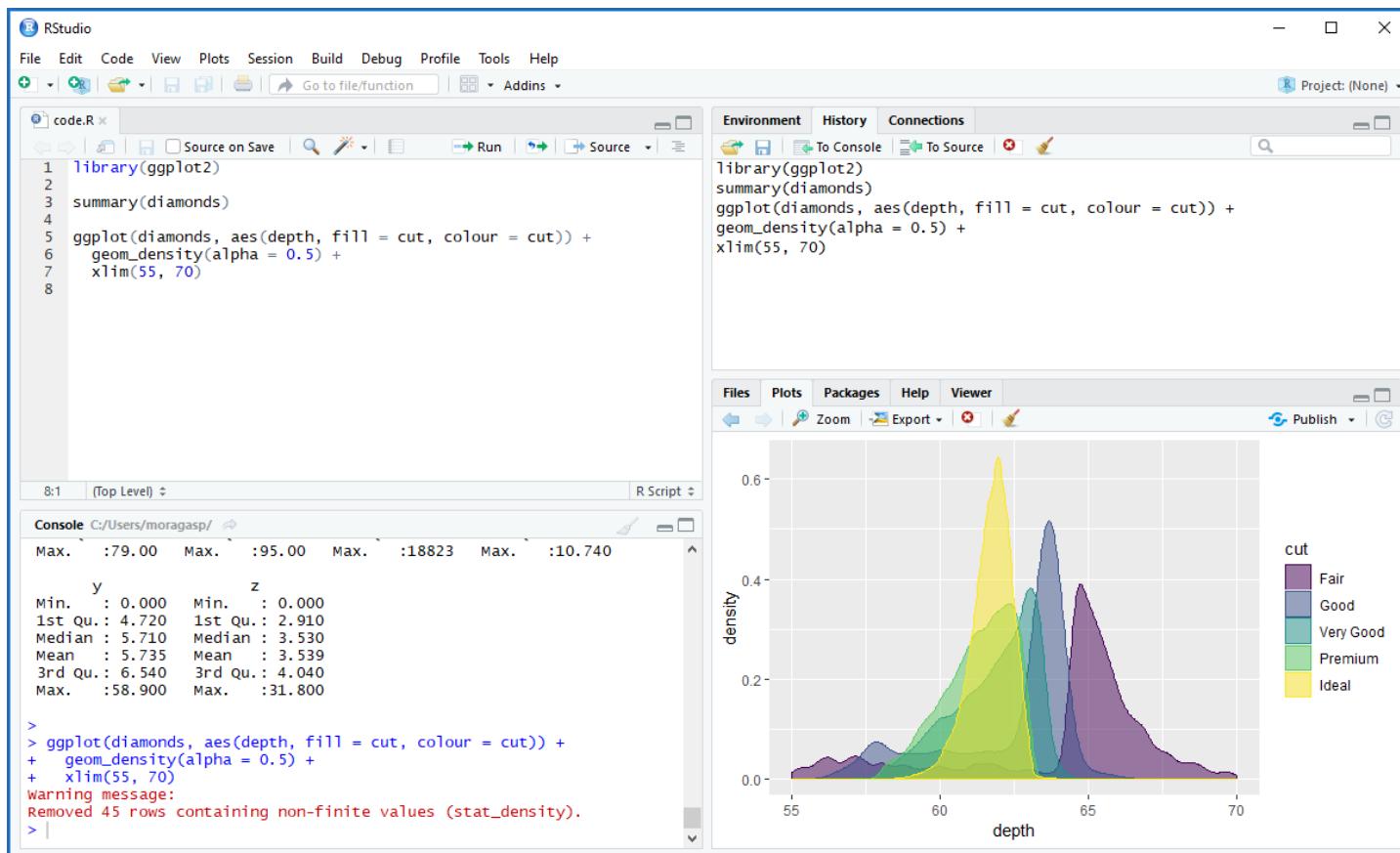


[https://en.wikipedia.org/wiki/R_\(programming_language\)](https://en.wikipedia.org/wiki/R_(programming_language))



- RStudio er et utviklingsmiljø (engelsk; integrated development environment (IDE)) for R.
 - <https://rstudio.com/>
- Tilbyr f.eks. bedre håndtering av mange simultant åpne filer, fargesetting av kode, håndtering av figurer, osv.

R/RStudio GUI



LÆRINGSMÅL I R FOR UKE 1

- En ren installasjon av R/Rstudio kommer med relativt lite funksjonalitet.
- Kraften til R ligger i muligheten til å legge til ekstrafunksjoner ved å installere pakker.

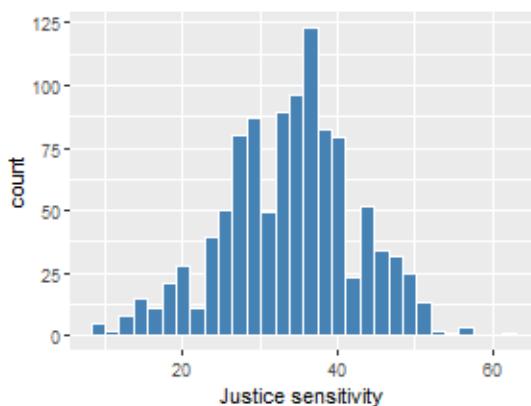
```
install.packages("tidyverse")
library(tidyverse)
```

- Denne uken skal vi i R se på:
 - Introduksjon til visualisering med ggplot2
 - Grunnleggende datahåndtering i R med dplyr
 - Utregning av beskrivende statistikker

HISTOGRAMMER GG PLOT

```
ggplot(data = <DATA>) +  
<GEO M_FUNCTION>(mapping = aes(<MAPPINGS>))
```

```
library(tidyverse)  
dt ← read_csv("https://www.sv.uio.no/psi/personer/vit/nikolaic/psy2014/rwa_dataset.dat")  
  
ggplot(data=dt, aes(x=JS)) +  
  geom_histogram(color="white", fill="steelblue") +  
  labs(x="Justice sensitivity")
```



ENKEL BESKRIVENDE STATISTIKK I R

```
# "c()" brukes til å lage en vektor  
X ← c(1,2,3,4,5,8,8,20,100)  
  
mean(X)
```

```
## [1] 16.77778
```

```
median(X)
```

```
## [1] 5
```

```
var(X)
```

```
## [1] 1006.194
```

```
sd(X)
```

```
## [1] 31.72057
```


TIDYVERSE

- The tidyverse er en samling av R-pakker designet for datavitenskap.
- Alle pakkene deler designfilosofi, grammatikk og datastrukturer.

```
# install.packages(tidyverse)
library(tidyverse)
```

TIBBLES OG DATA FRAMES

Tibbles er en moderne versjon av data-frames. Tenk på dem som en variabel som lagrer et excel-ark.

```
dt ← read_csv("https://www.sv.uio.no/psi/personer/vit/nikolaic/psy2014/rwa_dataset.dat")
dt
## # A tibble: 1,061 x 103
##   RWA1   RWA2   RWA3   RWA4   RWA5   RWA6   RWA7   RWA8   RWA9   RWA10  RWA11  RWA12  RWA13
##   <dbl> <dbl>
## 1     6     6     6     6     3     4     3     4     6     6     6     2     7
## 2     7     7     4     5     1     4     4     4     6     1     7     6     5
## 3     5     4     4     4     3     4     4     4     3     3     5     5     4
## 4     2     2     5     5     1     5     1     2     7     4     4     3     7
## 5     7     6     5     6     1     5     1     5     7     5     7     1     6
## 6     2     6     2     4     1     6     1     6     6     7     3     7     3
## 7     6     5     4     5     1     5     2     4     6     5     5     4     7
## 8     7     7     7     7     7     4     7     5     7     7     7     6     7
## 9     2     5     3     4     1     5     2     5     3     7     5     6     3
## 10    6     4     7     5     1     3     1     5     4     6     1     6     6
## # ... with 1,051 more rows, and 90 more variables: RWA14 <dbl>, RWA15 <dbl>,
## #   BFI1 <dbl>, BFI2 <dbl>, BFI3 <dbl>, BFI4 <dbl>, BFI5 <dbl>, BFI6 <dbl>,
## #   BFI7 <dbl>, BFI8 <dbl>, BFI9 <dbl>, BFI10 <dbl>, BFI11 <dbl>, BFI12 <dbl>,
## #   BFI13 <dbl>, BFI14 <dbl>, BFI15 <dbl>, BFI16 <dbl>, BFI17 <dbl>,
## #   BFI18 <dbl>, BFI19 <dbl>, BFI20 <dbl>, BFI21 <dbl>, BFI22 <dbl>,
## #   BFI23 <dbl>, BFI24 <dbl>, BFI25 <dbl>, BFI26 <dbl>, BFI27 <dbl>,
```

PIPE OPERATOR

Pipe operatoren, skrevet som `%>%`, er en viktig operator i `magrittr`-pakken. Den tar resultatet fra en funksjon og sender den som et argument til en annen funksjon.

```
dt %>% head(2)
## # A tibble: 2 x 103
##   RWA1   RWA2   RWA3   RWA4   RWA5   RWA6   RWA7   RWA8   RWA9   RWA10  RWA11  RWA12  RWA13
##   <dbl>  <dbl>
## 1     6     6     6     6     3     4     3     4     6     6     6     2     7
## 2     7     7     4     5     1     4     4     4     6     1     7     6     5
## # ... with 90 more variables: RWA14 <dbl>, RWA15 <dbl>, BFI1 <dbl>, BFI2 <dbl>,
## #   BFI3 <dbl>, BFI4 <dbl>, BFI5 <dbl>, BFI6 <dbl>, BFI7 <dbl>, BFI8 <dbl>,
## #   BFI9 <dbl>, BFI10 <dbl>, BFI11 <dbl>, BFI12 <dbl>, BFI13 <dbl>,
## #   BFI14 <dbl>, BFI15 <dbl>, BFI16 <dbl>, BFI17 <dbl>, BFI18 <dbl>,
## #   BFI19 <dbl>, BFI20 <dbl>, BFI21 <dbl>, BFI22 <dbl>, BFI23 <dbl>,
## #   BFI24 <dbl>, BFI25 <dbl>, BFI26 <dbl>, BFI27 <dbl>, BFI28 <dbl>,
## #   BFI29 <dbl>, BFI30 <dbl>, BFI31 <dbl>, BFI32 <dbl>, BFI33 <dbl>, ...
```

DPLYR PAKKEN

`dplyr` er en R-pakke som implementerer en grammatikk for datamanipulering. Følgende sett av funksjoner hjelper deg med å løse de vanligste datamanipulasjonsutfordringene.

- `filter()` velger rader basert på deres verdier.
- `select()` velger kolonner (variabler) basert på navn.
- `mutate()` legger til nye variabler som er funksjoner av eksisterende variabler.
- `arrange()` endrer rekkefølgen på radene.
- `summarise()` lager et sammendrag utifra et sett med verdier.

VELG RADER MED FILTER()

filter() velger rader basert på deres verdier.

```
dt %>% filter(age < 60)
## # A tibble: 264 x 103
##   RWA1   RWA2   RWA3   RWA4   RWA5   RWA6   RWA7   RWA8   RWA9   RWA10  RWA11  RWA12  RWA13
##   <dbl> <dbl>
## 1     6     6     6     6     3     4     3     4     6     6     6     2     7
## 2     7     7     4     5     1     4     4     4     6     1     7     6     5
## 3     6     2     5     4     2     4     7     4     6     1     6     1     6
## 4     7     7     4     7     1     7     6     1     6     1     6     4     7
## 5     5     1     1     7     1     6     1     7     4     7     6     5     7
## 6     5     6     4     5     1     7     1     4     4     6     5     2     4
## 7     5     3     5     5     4     6     3     5     6     3     6     1     4
## 8     1     7     1     7     1     7     1     1     1     7     6     7     5
## 9     5     6     6     5     1     6     1     5     6     2     5     5     4
## 10    5     5     2     6     2     6     6     5     4     1     5     2     5
## # ... with 254 more rows, and 90 more variables: RWA14 <dbl>, RWA15 <dbl>,
## #   BFI1 <dbl>, BFI2 <dbl>, BFI3 <dbl>, BFI4 <dbl>, BFI5 <dbl>, BFI6 <dbl>,
## #   BFI7 <dbl>, BFI8 <dbl>, BFI9 <dbl>, BFI10 <dbl>, BFI11 <dbl>, BFI12 <dbl>,
## #   BFI13 <dbl>, BFI14 <dbl>, BFI15 <dbl>, BFI16 <dbl>, BFI17 <dbl>,
## #   BFI18 <dbl>, BFI19 <dbl>, BFI20 <dbl>, BFI21 <dbl>, BFI22 <dbl>,
## #   BFI23 <dbl>, BFI24 <dbl>, BFI25 <dbl>, BFI26 <dbl>, BFI27 <dbl>,
## #   BFI28 <dbl>, BFI29 <dbl>, BFI30 <dbl>, BFI31 <dbl>, BFI32 <dbl>, ...
```

VELG KOLONNER MED SELECT()

select() velger kolonner (variabler) basert på navn.

```
dt %>% select(male, age)
## # A tibble: 1,061 x 2
##   male   age
##   <dbl> <dbl>
## 1     1    59
## 2     0    59
## 3     1    60
## 4     0    66
## 5     1    70
## 6     0    65
## 7     1    63
## 8     1    60
## 9     1    62
## 10    0    64
## # ... with 1,051 more rows
```

LAG NYE VARIABLER MED MUTATE()

mutate() legger til nye variabler som er funksjoner av eksisterende variabler.

```
dt %>% mutate(mean_js13 = (JS1+JS2+JS3)/3) %>% select(mean_js13)
## # A tibble: 1,061 x 1
##   mean_js13
##       <dbl>
## 1        4
## 2      3.67
## 3      3.33
## 4      2.67
## 5        2
## 6        1
## 7      2.67
## 8        2
## 9        2
## 10     2.33
## # ... with 1,051 more rows
```

SORTER RADER MED ARRANGE()

arrange() endrer rekkefølgen på radene.

```
dt %>% arrange (age) %>% select(age)
## # A tibble: 1,061 x 1
##       age
##   <dbl>
## 1     56
## 2     56
## 3     56
## 4     56
## 5     56
## 6     56
## 7     56
## 8     56
## 9     56
## 10    56
## # ... with 1,051 more rows
```

OPPSUMMER VARIABLER MED SUMMARIZE()

- `summarise()` lager et sammendrag utifra et sett med verdier.
- Funksjonen er spesielt kraftig i kombinasjon med `group_by()`.

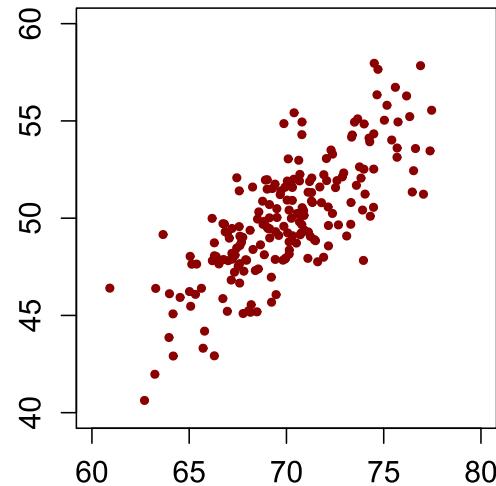
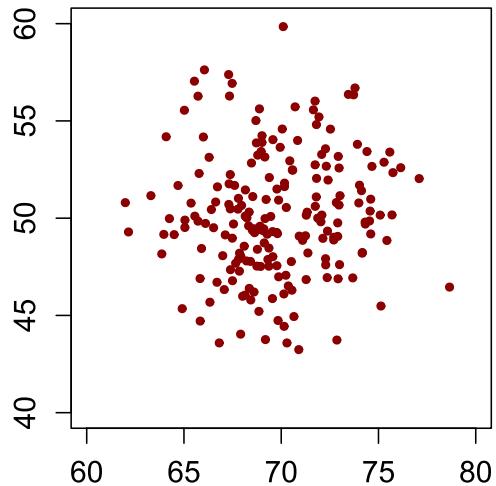
```
dt %>% summarize(mean_age = mean(age), sd_age = sd (age))
```

```
## # A tibble: 1 x 2
##   mean_age sd_age
##       <dbl>   <dbl>
## 1     63.2    4.44
```

```
dt %>% group_by(male) %>% summarize(mean_age = mean(age), sd_age = sd (age))
```

```
## # A tibble: 2 x 3
##   male mean_age sd_age
##   <dbl>     <dbl>   <dbl>
## 1     0      63.2    4.51
## 2     1      63.3    4.34
```

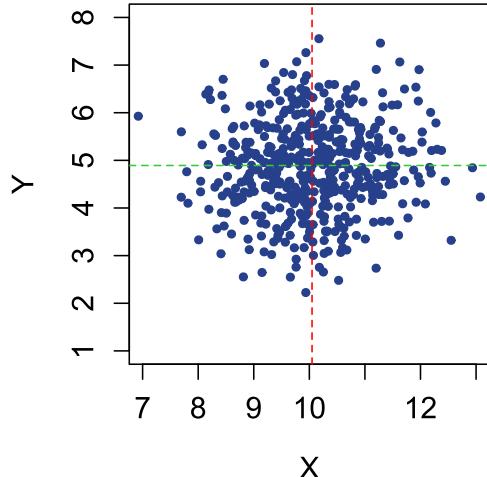

SPREDNINGSPLOT (SCATTER PLOTS)



Spredningsplot: En figur der hvert par med verdier på to variabler plottes som et punkt i planet.

- Uavhengig variable utgjør typisk X-verdien.
- Avhengig variabel utgjør typisk Y-verdien.

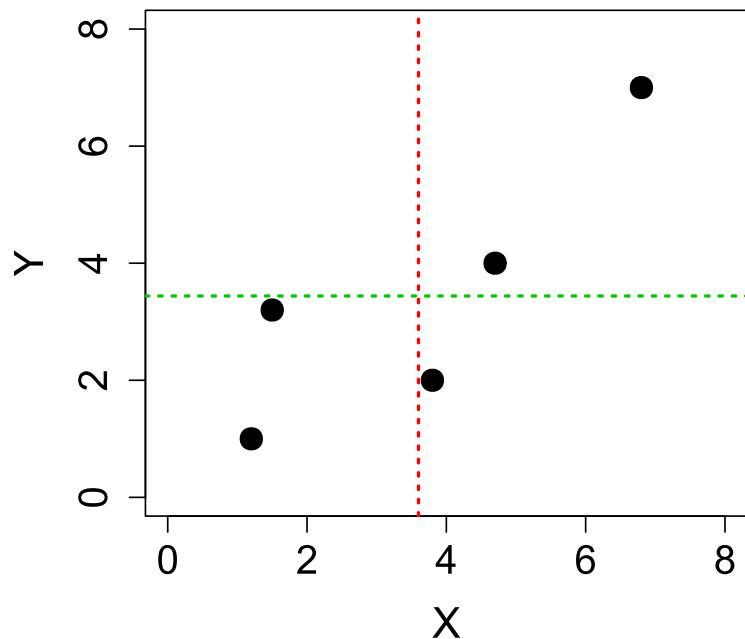
KOVARIANS



$$cov(x, y) = s_{xy} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

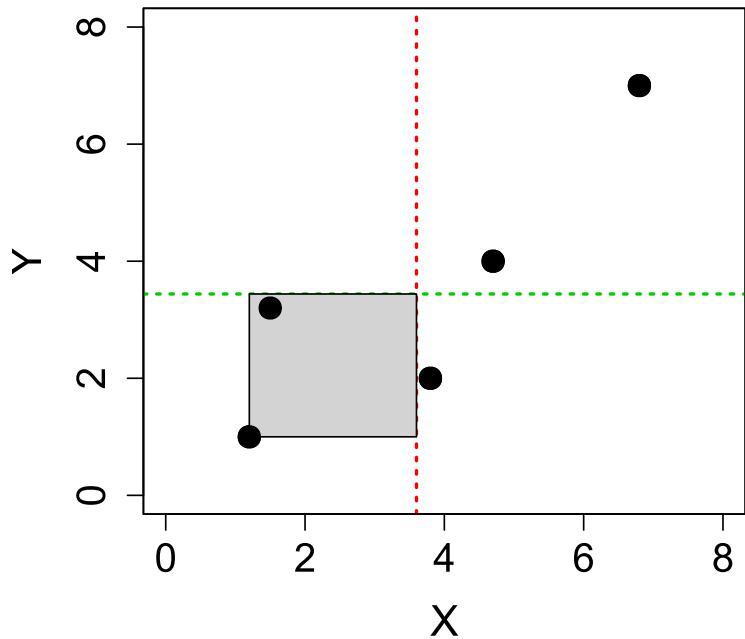
Kovarians er en statistikk der grad av samvariasjon mellom to variabler kvantifiseres.

UTREGNING AV KOVARIANS



BEREGNING AV KOVARIANS

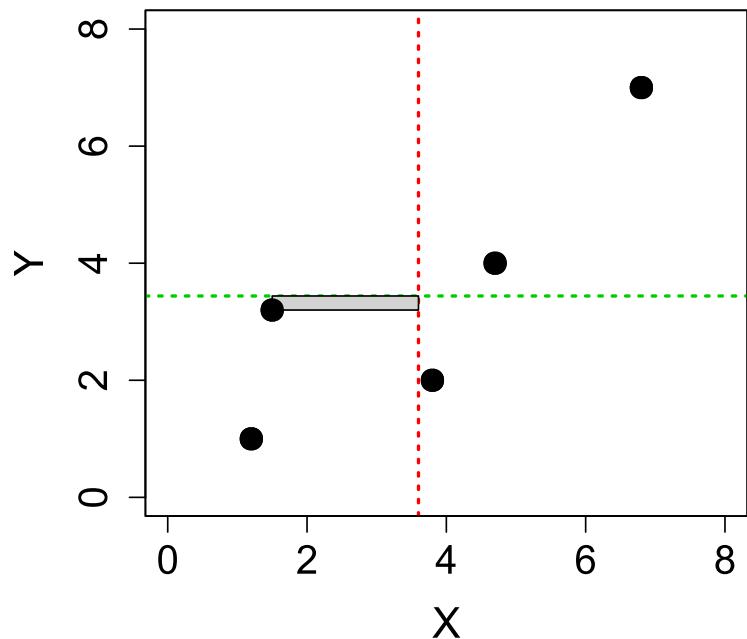
$$(X_1 - \bar{X})(Y_1 - \bar{Y}) = (1.2 - 3.6)(1 - 3.44) = (-2.4)(-2.44) = 5.86$$



$$s_{xy} = \frac{1}{4}(5.86)$$

BEREGNING AV KOVARIANS

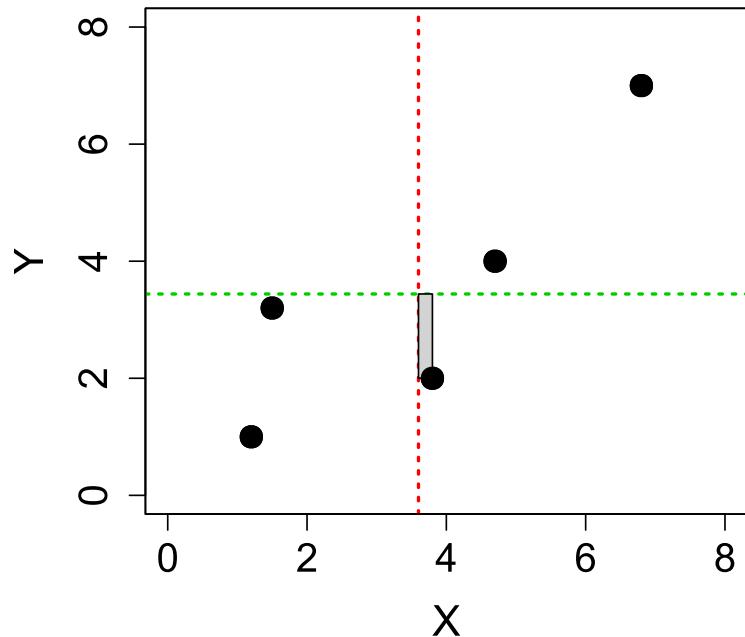
$$(X_2 - \bar{X})(Y_2 - \bar{Y}) = (1.5 - 3.6)(3.2 - 3.44) = (-2.1)(-0.24) = 0.5$$



$$s_{xy} = \frac{1}{4}(5.86 + 0.15$$

BEREGNING AV KOVARIANS

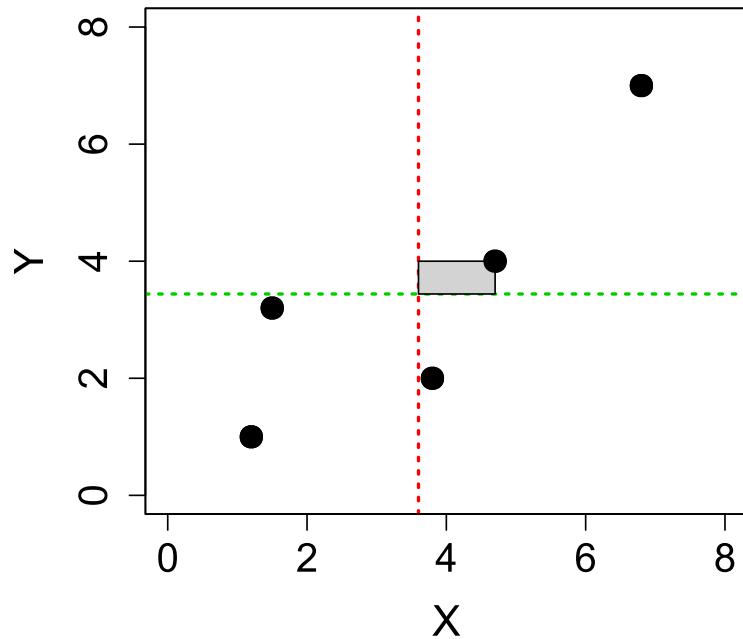
$$(X_3 - \bar{X})(Y_3 - \bar{Y}) = (3.8 - 3.6)(2 - 3.44) = (0.2)(-1.44) = -0.29$$



$$s_{xy} = \frac{1}{4}(5.86 + 0.15 + (-0.29))$$

BEREGNING AV KOVARIANS

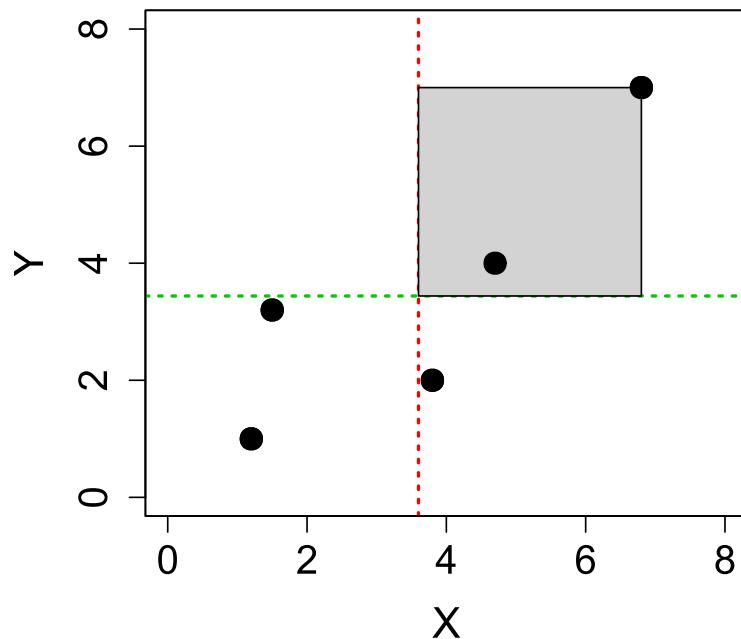
$$(X_4 - \bar{X})(Y_4 - \bar{Y}) = (4.7 - 3.6)(4 - 3.44) = (1.1)(0.56) = 0.62$$



$$s_{xy} = \frac{1}{4}(5.86 + 0.15 + (-0.29) + 0.62)$$

BEREGNING AV KOVARIANS

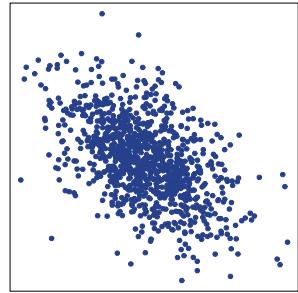
$$(X_5 - \bar{X})(Y_5 - \bar{Y}) = (6.8 - 3.6)(7 - 3.44) = (3.2)(3.56) = 11.39$$



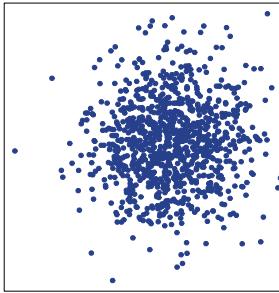
$$s_{xy} = \frac{1}{4}(5.86 + 0.15 + (-0.29) + 0.62 + 11.39) = 18.08$$

PEARSON CORRELATION

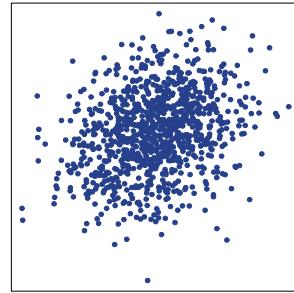
Pearson korrelasjon = -0.5



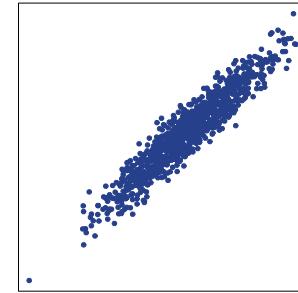
Pearson korrelasjon = 0.0



Pearson korrelasjon = 0.3



Pearson korrelasjon = 0.95



$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

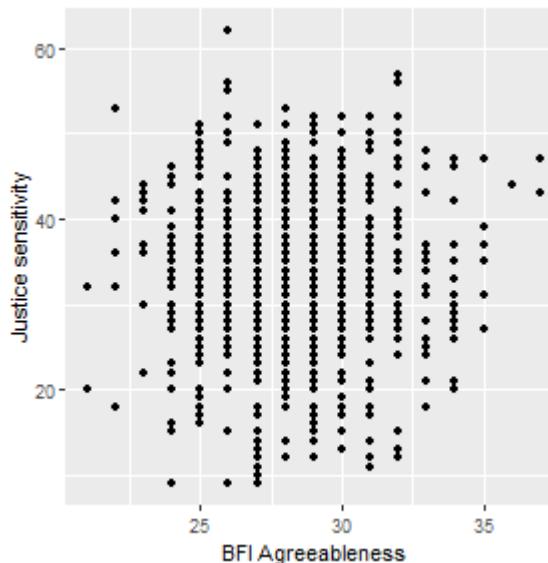
Pearson-korrelasjon:

- utgjør en *standardisering* av kovariansen.
- er definert som kovariansen delt på den største verdien denne kovariansen kan ta (produktet av variablenees standardavvik).
- er den vanligste statistikken brukt for å bedømme styrken av forholdet mellom to variabler.

SCATTER PLOT OG KORRELASJON I R

```
library(tidyverse)
dt ← read_csv("https://www.sv.uio.no/psi/personer/vit/nikolaic/psy2014/rwa_dataset.dat")

ggplot(dt, aes(x=BFI_A, y=JS))+  geom_point()+
  labs(x="BFI Agreeableness", y="Justice sensitivity")
```



```
dt %>% select(JS,BFI_A) %>% cor() %>% round(2)
```

```
##          JS BFI_A
## JS     1.00  0.06
## BFI_A  0.06  1.00
```

FORKLART VARIANS

- **Forklart varians (r^2): Andelen varians to variabler har til felles.**
 - Et tall mellom 0 og 1 (siden korrelasjonen er mellom -1 og 1).
- Feks. Anta at korrelasjonen mellom inntekt og empati er -0.4. Hvor mye av variabiliteten i empati kan du gjøre rede for dersom du kjenner inntekten til alle i utvalget?
 - $(-0.4)^2 = 0.16$, (or 16%)
- **Korrelasjon er derfor et mål på effektstørrelse.**
 - $r = 0.1$ (liten effekt): 1% av variansen er forklart.
 - $r = 0.3$ (middels effekt): 9% av variansen er forklart.
 - $r = 0.5$ (stor effekt): 25% av variansen er forklart.