



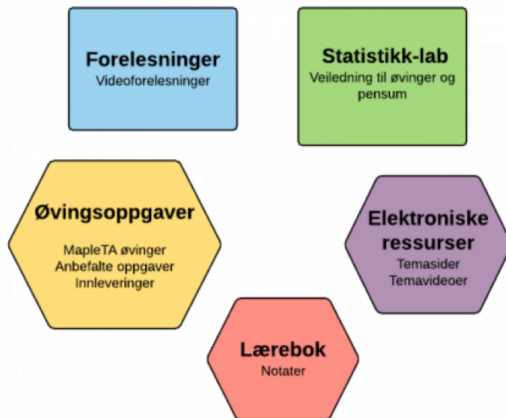
Velkommen til TMA4240

Planen for idag

- Praktisk info
- Hva er statistikk
- Hva skal dere lære i dette kurset
- Deskriptiv statistikk

TMA4240: Tilgjengelig resurser

- Faglærer: Sara Martino og Torstein Fjeldstad
- Hjemmeside:
<https://wiki.math.ntnu.no/tma4240/2018h/start>



- To Parallell:
 - Mandag-Fredag 10:15-12:00
 - Tirsdag 14:15-16:00, Torsdag 16:15-18:00
- Du kan velge hvilken forelesning du ønsker
- Begge parallel har samme pensum og de er delt "verticalt" mellom to foreleser
 - 1 Blok: Sara Martino
 - 2 Block: Torstein Fjeldstad

**Elektronisk
øving**

**Anbefalt
øving**

**Skriftlig
øving**

Statistikklab

Krav til godkjent øvingsopplegg

- Elektronisk (Maple TA) øving
 - Minst 4 godkjente øvinger fra Blokk 1
 - Minst 4 godkjente øvinger fra Blokk 2
- Skriftlige innleveringer
 - 3 godkjente innleveringer, hvorav minst 1 må være fra Blokk 1 og minst 1 fra Blokk 2

Her kan du samarbeide med andre og få hjelp av studentassistenter

Dag	Tid	Rom
Mandag	14:15-18:00	S4
Torsdag	12:15-16:00	S1
Fredag	12:15-16:00	S4 *

(*) I uke 36 er det flyttet i S1

- **Walpole, Myers, Myers and Ye (2012): Probability and Statistics for engineers (9. utgave), Pearson / Prentice Hall.**
 - Det finnes en egen utgave for NTNU (ISBN 978-1-78726-081-8) som selges på Akademika.
 - Det finnes en NTNU-versjon av 9. utgave med ISBN 978-1-78448-014-1 som også kan brukes. Originalversjonen (internasjonal versjon) av 9. utgave kan også brukes.
 - Det finnes en utgave med **blå skrift** (ISBN 10:1-292-02392-9. ISBN 13: 978-1-292-02392-2), som **vi fraråder** da flere har gitt tilbakemelding om at den er vanskelig å lese og å finne frem i.
- **Tabeller of formel i statistikk. Akademika**

- Temasider
- Temavideo

- Vi trenger 2/3 studenter som kan være i referansegruppe
 - 2/3 møter i løpet av semester
 - skrive en kort rapport til slutt

Hva er Statistikk??

Hva er Statistikk??

Statistikk er en gren av matematikk som tar for seg analyse av tallfestede dataobservasjoner med den hensikt å kunne trekke konklusjoner basert på disse dataene

Statistikk er en verktøy for å ta beslutnings under usikkerhet og variasjon.

Statistikk handler om å tallfeste vår usikkerhet

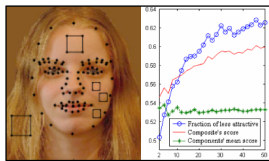
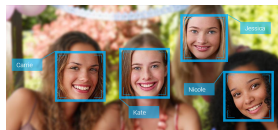
Recommender systems



- Basert på hvordan andre har vurdert en product vil amazon vite hvordan *du* ville vurdere en produkt du ikke har vurdert
- I 2006 ga Netflix tilgang til 100480507 vurdering gitt av 480189 brukere til 17770 filmer

Statistikk handler om "Big Data"...

Face recognition



- Finn ansiktet i en bilde
- finn ut om ansiktet er for eksempel i dårlig belysning eller i profil
- velge ut unike egenskaper i ansiktet
- sammenlign de unike egenskapene til det ansiktet til alle menneskene du allerede kjenner

Statistikk handler om lite data...

30 av 100 ble friske ved bruk av medisin A og 20 av 50 ble friske ved bruk av medisin B. Er medisin A bedre enn medisin B?

Statistikk handler om lite data...

30 av 100 ble friske ved bruk av medisin A og 20 av 50 ble friske ved bruk av medisin B. Er medisin A bedre enn medisin B?

- 40% er høyere enn 30%, men er det nok?
- Det kan være svært kostbart å trekke feil konklusjon
- Statistikk svarer på spørsmålet: hvor sikker er vi på at medisin B er bedre enn medisin A?

Statistikk handler om å tallfeste vår usikkerhet

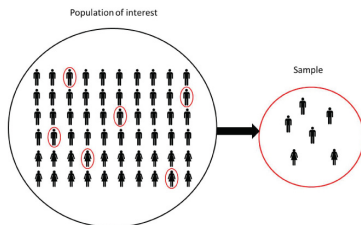
Vi vil vite hvor høyt nordmenn er i gjennomsnit.

- Veldig lett å svare: mål alle og ta gjennomsnit!!!
- ...men det er veldig komplisert å måle alle!!

Statistikk handler om å tallfeste vår usikkerhet

Vi vil vite hvor høyt nordmenn er i gjennomsnitt.

- Veldig lett å svare: mål alle og ta gjennomsnitt!!!
- ...men det er veldig komplisert å måle alle!!
- Vi må måle noe (et utvalg) og bruke vår resultat til å gi et svar for hele populasjon
- Resultate blir da usikkert
- Det beste vi kan gjør er å tallfeste hvor mye usikkerhet er det rundt vår svar



De finnes mange forskjellige data...

Kontinuerlig data Høyde, vekt, temperatur

Diskret data Antall studenter som tar TMA4240, antall søkere til NTNU, antall øyne på en terning

Kategorisk data Syk/ikke-syk, øyefarge, stryk/bestått

..og mange forskjellige problemer

- Hva er årlig tilsig i en vassdrag hvor vi ikke har målinger?
- Er det en trend i temperatur?
- Lever røykere kortere enn ikke-røykere?
- Hva blir været i morra?



Map of the Vansjø-Hobøl Catchment, Norway. Image: Skarbøvik and Bechmann (2010). Bioforsk Report vol. 5

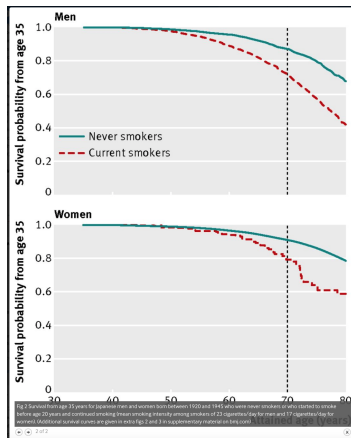
..og mange forskjellige problemer

- Hva er årlig tilsig i en vassdrag hvor vi ikke har målinger?
- Er det en trend i temperatur?
- Lever røykere kortere enn ikke-røykere?
- Hva blir været i morra?



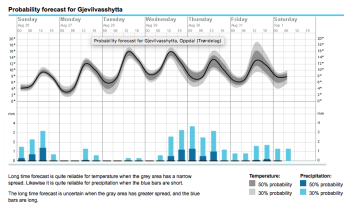
..og mange forskjellige problemer

- Hva er årlig tilsig i en vassdrag hvor vi ikke har målinger?
- Er det en trend i temperatur?
- **Lever røykere kortere enn ikke-røykere?**
- Hva blir været i morra?

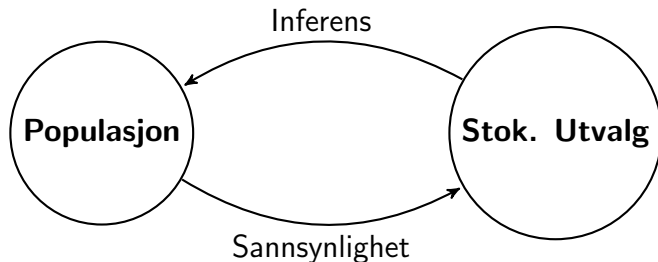


..og mange forskjellige problemer

- Hva er årlig tilsig i en vassdrag hvor vi ikke har målinger?
- Er det en trend i temperatur?
- Lever røykere kortere enn ikke-røykere?
- Hva blir været i morra?



Hva skal vi lære i dette kurset?



- Deskriptiv Statistikk, Kap 1 (Idag)
 - Presentere og beskrive data i tall, tabeller, figurer
- Sannsynlighet, Kap 2-7 (Fra neste gang)
 - regne men sannsynlighet
- Statistikk inferens, Kap 8-11 (Senere)
 - trekke slutninger om en hel populasjon fra observasjoner på et utvalg

Sannsynlighetseksempel Du har en rettferdig mynt (like sannsynlighet for hoder eller haler). Hva er sannsynligheten for 60 eller flere hoder hvis du kaster mynt 100 ganger?

Statistikk eksempel Du har en ukjent mynt. Du kaster du det 100 ganger og teller 60 hoder. Er mynten rettferdig?

Sannsynlighetseksempel Du har en rettferdig mynt (like sannsynlighet for hoder eller haler). Hva er sannsynligheten for 60 eller flere hoder hvis du kaster mynt 100 ganger?

Det er bare ett svar (ca. 0.028444), og vi skal lære å beregne det.

Statistikk eksempel Du har en ukjent mynt. Du kaster du det 100 ganger og teller 60 hoder. Er mynten rettferdig?

Sannsynlighetseksempel Du har en rettferdig mynt (like sannsynlighet for hoder eller haler). Hva er sannsynligheten for 60 eller flere hoder hvis du kaster mynt 100 ganger?

Det er bare ett svar (ca. 0.028444), og vi skal lære å beregne det.

Statistikk eksempel Du har en ukjent mynt. Du kaster du det 100 ganger og teller 60 hoder. Er mynten rettferdig? Det er mange måter å gå videre. Faktisk kan forskjellige statistikere trekke forskjellige konklusjoner.

Sannsynlighetseksempel Tilfeldige prosessen fullt kjent (sannsynlighet for hodene = .5).

Målet er å finne sannsynligheten for et bestemt utfall (minst 60 hoder).

Statistikk eksempel Resultatet kjent (60 hoder) og målet er å belyse den ukjente tilfeldige prosessen (sannsynligheten for hoder).

Descriptive Statistikk (Kap. 1)

Kvinner: 170.18 170.18 170.18 152.4 172.72 162.56 175.26
180.34 170.18 170.18 167.64 160.02 170.18 157.48 167.64
177.8 170.18 154.94 172.72 170.18 172.72 162.56 175.26
170.18 177.8 182.88 154.94 170.18 175.26 172.72 175.26
182.88 167.64 170.18 167.64 170.18 175.26 162.56 162.56
160.02 172.72 167.64 165.1 152.4 177.8 165.1 172.72 167.64
154.94 165.1

Hvor begynner man??

Vanskelig å se på en rekke tall.

Men vi kan

- Lage plott og tabeller som viser hvordan de observerte verdiene er spredt i forhold til hverandre
- Regne ut tall som beskriver sentrene til de observerte verdiene
- Regne ut tall som beskriver spredningene til de observerte verdiene

Dette er metoder for å beskrive de observerte verdiene og bruk av disse kalles **deskriptiv statistikk**

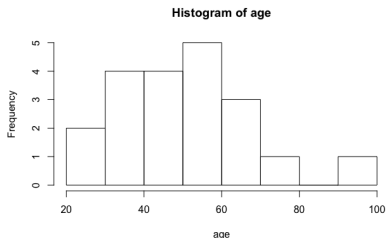
Histogram

Age: 36, 25, 38, 46, 55, 68, 72, 55, 36, 38, 67, 45, 22, 48, 91, 46, 52, 61, 58, 55

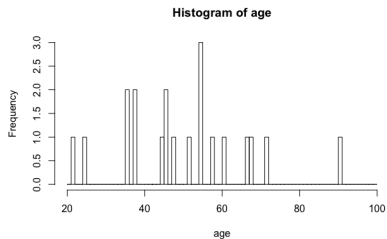
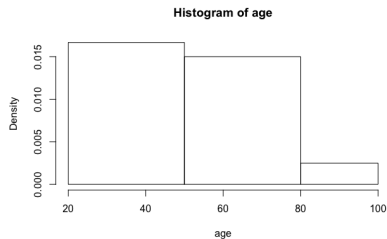
Histogram

Age: 36, 25 ,38 ,46, 55, 68, 72, 55 ,36 ,38, 67, 45, 22, 48, 91, 46, 52, 61, 58, 55

Bin	Freq	
20-30	2	25,22
30-40	4	36,38,36,38
40-50	4	46,45,48,46
50-60	5	55,55,52,58,55
60-70	3	68,67,61
70-80	1	72
80-90	0	-
90-100	1	91



Viktig med bin størrelse



Definition (Empirisk Middelerdi)

Den empirisk middelerdi til observasjonene $x_1, x_2, \dots, x_n$ er gitt ved

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

Definition (Empirisk Middelvei)

Den empirisk middelvei til observasjonene $x_1, x_2, \dots, x_n$ er gitt ved

$$\bar{x} = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

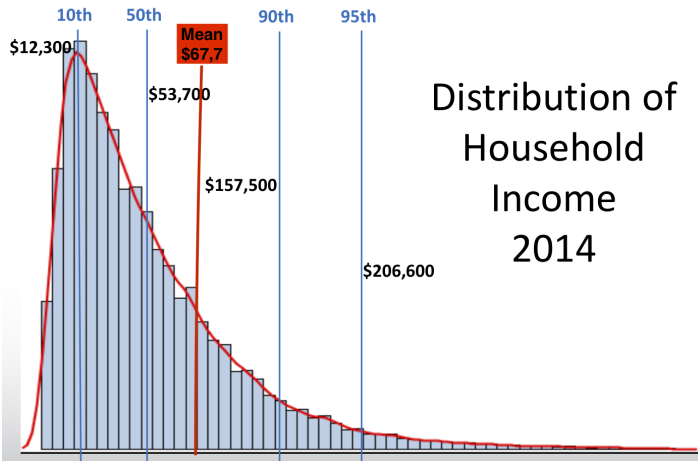
Definition (Empirisk Median)

Den empirisk median til observasjonene $x_1, x_2, \dots, x_n$ er gitt ved

$$\tilde{x} = \begin{cases} x_{(n+1)/2}, & \text{for } n \text{ odde} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & \text{for } n \text{ like} \end{cases}$$

hvor $x_{(1)}, \dots, x_{(n)}$ indikerer sorterte verdiene ($x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$)

Empirisk gjennosnitt vs Empirisk median



Source: U.S. Census Bureau, Current Population Survey, 2015 Annual Social and Economic Supplement.

United States
Census
Bureau

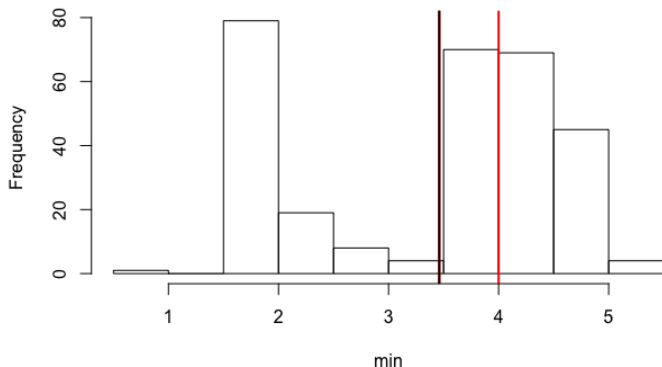
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Empirisk gjennomsnitt vs Empirisk median

Emp. Gjennomsnitt

Emp. Median

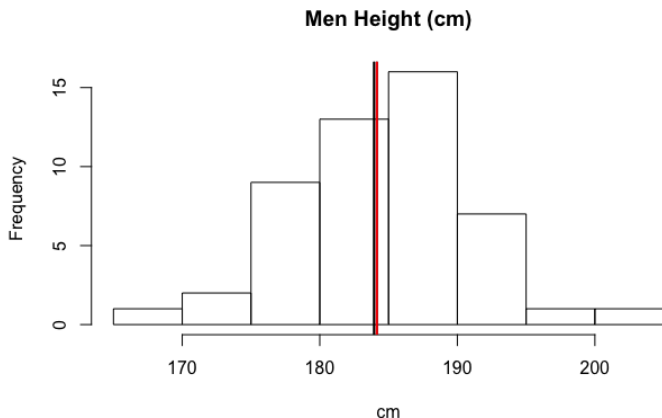
Duration of the eruption



Empirisk gjennomsnitt vs Empirisk median

Emp. Gjennomsnitt

Emp. Median



Empirisk gjennomsnitt vs Empirisk median

- Empirisk gjennomsnitt er veldig påvirket av ekstreme verdier
- Empirisk median er mer robust
- Empirisk median mer brukt med skjev fordelte data

Sammenligning av spredning

Det mest vanlige målet på variabilitet er empirisk varians

Definition (Empirisk varians)

Den empirisk varians til observasjonene $x_1, x_2, \dots, x_n$ er gitt ved

$$s^2 = \frac{1}{n-1} [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2] = \frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2$$

Empirisk varians kan være vanskelig å tolke fordi den er uttrykt i kvadratiske enheter. Empirisk standardavvik er uttrykt i samme enheter som dataene.

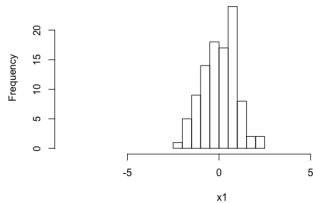
Definition (Empirisk Standardavvik)

Den empirisk varians til observasjonene $x_1, x_2, \dots, x_n$ er gitt ved

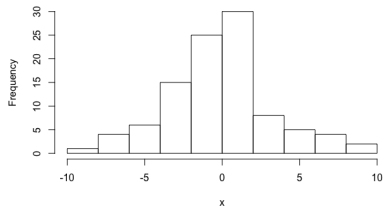
$$s = \sqrt{s^2} = \sqrt{\frac{1}{(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2}$$

Spredning

mean=0, sd=1



mean=0, sd=3



Mileage per Gallon

Manual 21.0 21.0 22.8 32.4 30.4 33.9 27.3 26.0 30.4 15.8 19.7 15.0
21.4

Automatic 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4
10.4 14.7 21.5 15.5 15.2 13.3 19.2

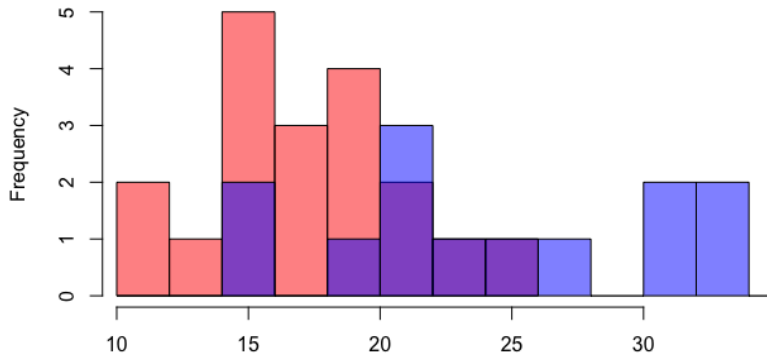
Mileage per Gallon

Manual 21.0 21.0 22.8 32.4 30.4 33.9 27.3 26.0 30.4 15.8 19.7 15.0
21.4

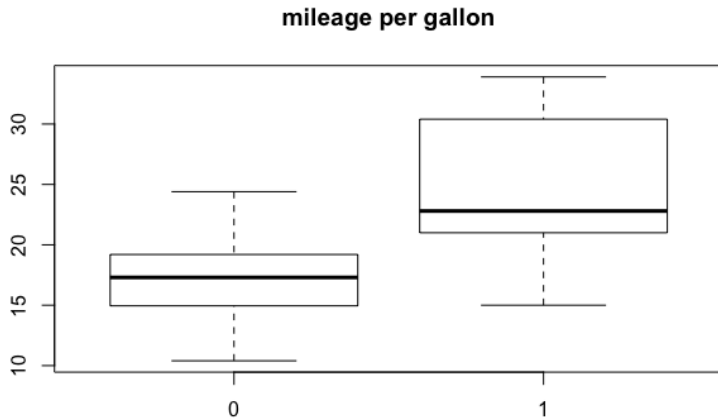
Automatic 21.4 18.7 18.1 14.3 24.4 22.8 19.2 17.8 16.4 17.3 15.2 10.4
10.4 14.7 21.5 15.5 15.2 13.3 19.2

	Emp. Gjenn. mil/gall	Emp. var (mil/gall) ²	Emp sd mil/gall	Min mil/gall	Max mil/gall
Automatic	17.1	14.6	3.8	10.4	24.4
Manual	24.3	38.02	6.1	15.0	33.9

Mileage per Gallon



Mileage per Gallon



Deskriptiv statistikk er en viktig startpunkt for å utforske de observerte verdiene.

Hjelper å skaffe seg en “bilde” av datane.

Men:

- Ingen bruk av sannsynlighet. Vi har ingen tall på hvor sikker vi er på at menn er høyere enn kvinner.
- Veldig subjektivt om man anser det for å være en forskjell eller ikke

Deskriptiv statistikk er en viktig startpunkt for å utforske de observerte verdiene.

Hjelper å skaffe seg en “bilde” av datane.

Men:

- Ingen bruk av sannsynlighet. Vi har ingen tall på hvor sikker vi er på at menn er høyere enn kvinner.
- Veldig subjektivt om man anser det for å være en forskjell eller ikke

Neste gang begynner vi reisen mot statistisk inferens hvor vi gjennom sannsynlighetsteori kan tallfeste vår sikkerhet til konklusjonene vi gjør