Beskrivende statistikk

Statistikk gjelder innsamling og behandling av data. Dataene er representert med tall. Statistikk er en gren av matematikken og det benyttes mye beregninger, formler, algebra osv. Navnet er avledet av «stat» fordi det opprinnelig typisk var *stater* som samlet inn data om sine borgere og samfunnsforhold.

Eksempel 1: Helt siden 1705 har alle rekrutter fått målt sin høyde og vekt på sesjon. Dette gir en meget stor mengde med tall for svært mange årskull. Målingene ble bl.a. brukt til å sy uniformer i riktig størrelser.

Ofte blir det svært mange tall og det er vanskelig (umulig?) å få noen oversikt over dem. I 2010 var det f.eks. over 26 000 menn på sesjon. Dermed blir ikke målingene særlig brukbare til noe.

I den **beskrivende statistikken** forsøker man å beskrive alle de mange dataene med færre tall. Det skal gi bedre oversikt og gjøre målingene nyttigere. Man kan

1. ***Gruppere dataene***. Da samles dataene i grupper. På den måten blir det færre tall og lettere å få oversikt. Samtidig mister man informasjon, da mange data med forskjellig verdi havner i samme gruppe. Vi vet at de hører til gruppen, men ikke hva hver enkelt verdi er.

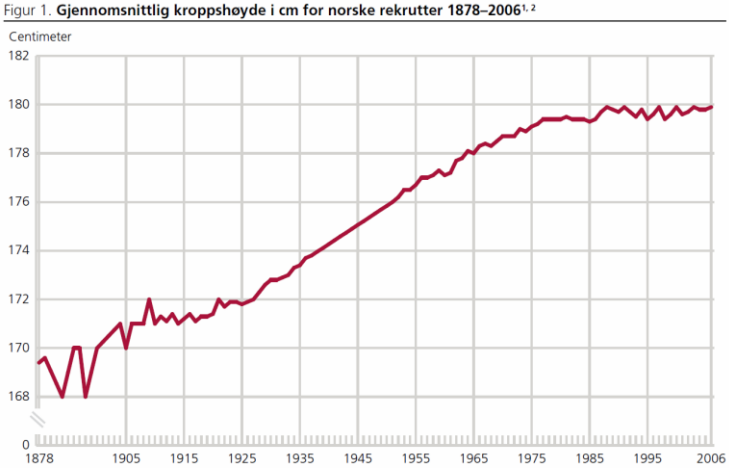
   Eksempel 2: Målingene av rekrutters høyde (menn) i 2008, fordelte seg slik gruppert:

Rekrutthøyder 2008		
Høyde i cm	Prosent-andel per klasse	Relativ frekvens
[160,165⟩	1,2	0,012
[165,170⟩	4,9	0,049
[170,175⟩	15,4	0,154
[175,180⟩	26,8	0,268
[180,185⟩	28,2	0,282
[185,190⟩	16,5	0,165
[190,195⟩	5,7	0,057
[195,200⟩	1,3	0,013
Sum	100	1

   Legg merke til at dette er *kontinuerlige* data (= uten sprang). Derfor er gruppene angitt som «halvåpne intervaller». I den første gruppen er f.eks. de med som målte nøyaktig 160 cm, men de som målte 165 cm er ikke med. Intervallet sies å være «lukket» i nedre ende og «åpent» i øvre ende der vi ikke kan si hvilket tall som er det største.
   Ordet *frekvens* betyr «hvor ofte forekom dette». At den er *relativ* innebærer at antallet er delt med totalantallet. Det er ganske uvanlig at totalantallet ikke er angitt, da det sier noe om hvor sikre tallene er. Hvis du har tellet opp 100 rekrutter er de usikre, med 26.000 rekrutter er de mye sikrere («de store talls lov»).

Note: Slik ser utviklingen ut for høyden 1878 til 2006, iflg. SSB:

Figur 1. **Gjennomsnittlig kroppshøyde i cm for norske rekrutter 1878–2006**[1, 2]



Samtidig øker BMI ☺

2. ***Beregne sentralmål***. Det innebærer å angi ett, eneste tall som på en eller annen måte representerer «sentrum» i dataene. Det er flere vanlige:

    a. *Gjennomsnitt*. Gjennomsnittet er summen av alle dataene delt med antallet. Dette er et meget vanlig sentralmål.

    Eksempel 3: Dataene for mannlige rekrutter viser at de er blitt høyere gjennom årene. I 1761 var f.eks. gjennomsnittshøyden ca 170 cm, i 2006 var den ca 180. Høyden har stagnert de siste 10-15 årene men rekruttene er blitt ca 3 kg tyngre i den samme perioden.

    b. *Median*. Medianen er den midterste verdien når dataene sorteres etter størrelse. (Hvis antallet verdier er et partall, er medianen verdien midt mellom de to midterste. I grupperte data finnes medianen ved interpolasjon i den gruppen som inneholder den midterste.)

    Eksempel 4: Jeg har ikke klart å finne medianhøyden for rekruttene, men utfra tabellen i eksempel 2 kan man se at medianen må være like over 180 cm. Interpolasjon i gruppen [180,185> gir median = 180,5.

    c. *Typetall*. Typetallet er den verdien som det finnes flest av blant dataene. (Hvis flere er like, oppgis alle.) Typetall passer best for data som måles diskontinuerlig (i sprang). Når dataene er kontinuerlige, er det egentlig ikke to *helt* like verdier. Rekruttenes høyde er jo egentlig kontinuerlige data, men når de måles til nærmeste centimeter, er de allikevel diskontinuerlige. Noen ganger angis den største gruppen i grupperte data som typetall, men det er nokså tvilsom praksis.

    Eksempel 5: Den største gruppen for rekruttenes høyde hentet fra tabellen i eksempel 2, er [180,185>. Noen vil da angi denne gruppen som typetall. Det kan virke rimelig at siden det er den mest tallrike gruppen, så befinner typetallet (den verdien som det er flest av) inne i denne gruppen. Det kan vi imidlertid egentlig ikke vite. For å finne det riktige typetallet, må vi se alle dataene.

3. ***Beregne spredningsmål***. Slike skal gi et inntrykk av hvor spredt dataene er. Det er flere slike også:
    a. *Variasjonsbredde*. Variasjonsbredden er forskjellen mellom største og minste verdi.

    Eksempel 6: Jeg har ikke grunnlagstallene tilgjengelig, men antar vi at laveste rekrutt i

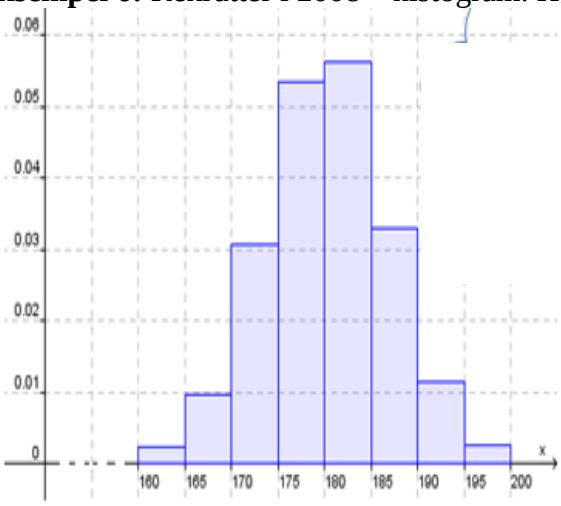
2008 ble målt til 161 cm og høyeste til 199 cm, så er variasjonsbredden 199 – 161 = 38 cm.

b. *Persentiler, kvartiler*. Disse målene tilsvarer medianen ved at dataene først må sorteres etter størrelse. Medianen er den verdien som nås 50% oppover mot toppen. Det innebærer da at 50% av verdiene er mindre enn medianen, mens 50% er større. Første kvartil er tilsvarende den verdien som har 25% som er lavere og 75% som er høyere. Tredje kvartil har 75% av verdiene lavere. Persentiler er andre prosenter som f.eks. 10%, 33,3% osv.

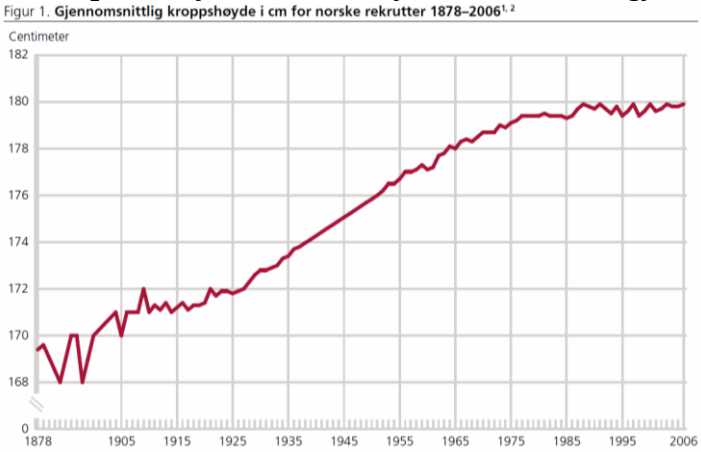
c. *Varians, standardavvik*. Dette er de beste spredningsmålene, men også de som er tyngst å beregne og vanskeligst å forstå. De angir en form for gjennomsnittlig avvik fra midten, altså hvor langt fra gjennomsnittsverdien (sentralmålet) verdiene ligger i gjennomsnitt. Hvis standardavviket er stort, er det mange verdier som er langt større og langt mindre enn gjennomsnittsverdien. Hvis standardavviket er lite, så ligger verdiene tett samlet omkring gjennomsnittsverdien. Variansen er kvadratet av standardavviket (Varians = Standardavvik^2) – et triks som brukes for at en verdi som er større en gjennomsnittsverdien og en som er mindre, skal telle likt (og positivt begge to). Vi ser ikke nærmere på dette her.

4. ***Lage grafer***. Med grafer mener vi tegninger av forskjellig slag. De gir ofte fin oversikt. Her er noen eksempler for rekruttene som er brukt ovenfor:

Eksempel 6: Rekrutter i 2008 – histogram. Høyden på søylene angir andelen i hver gruppe:



Eksempel 8: Gjennomsnittshøyden for rekrutter gjennom mange år:



Grafen viser at de mannlige rekruttene er blitt høyere etter hvert. I 1878 var f.eks.

gjennomsnittshøyden like under 170 cm, i 2006 var den ca 180. Som det fremgår har høyden stagnert de siste 10-15 årene.

Ovenfor er det forklart hvordan man kan få oversikt over store datamengder. Det er på tide å gjøre noen oppgaver!

Oppgave 1: Vi har samlet inn et utvalg blåskjell og målt lengden på skallet deres Vi fant følgende (3-9)

Lengde
5,8
6,5
5,6
4,9
5,7
5,1
6,3
5,0
4,0
4,2
7,3
5,0
5,5
6,1
5,6
5,9
6,1
6,3
5,6
6,4
5,0
5,4
5,4
5,2
6,8
7,2
4,9
4,3
5,6
6,4

a. Beregn gjennomsnittet
b. Sorter dataene og gi hver måling et nummer fra 1 til 30
c. Finn medianen
d. Finn variasjonsbredden
e. Finn typetallet
f. Avrund alle tallene til nærmeste centimeter
g. Finn typetallet for de avrundede dataene
h. Grupper de avrundede dataene i grupper for hver hele centimeter
i. Tegn histogram for de grupperte dataene
j. Plott inn de opprinnelige, *usorterte* dataene i et diagram (tegn ett punkt for hver måling – bruk målingsnummeret bortover og centimeter oppover)

(Jeg tenker å vise hvordan alt dette kan gjøres i et regneark på datamaskin.)