Jan Erik Kristiansen

Mot normalt: Om gjennomsnitt

Ved siden av prosenter er nok gjennomsnittet det statistiske begrepet de fleste kjenner til og føler seg fortrolige med. I dagligtale omtales gjennomsnittet ofte som «det vanlig(st)e» eller «det typiske». Andre synonymer er «middeltall», «tverrsnitt» eller «det normale». Å være gjennomsnittlig betyr som oftest noe i retning av å være som de fleste andre.

Gjennomsnitt er lett å beregne …

Gjennomsnittet brukes i en rekke beskrivelser av samfunnet: Gjennomsnittsinntekten for husholdninger er omtrent 366 000 kroner, gjennomsnittlig fødealder er vel 30 år, gjennomsnittsalderen for lærere i videregående skole er 49 år, gjennomsnittstemperaturen i Oslo i mai er 10,2 grader, gjennomsnittlig arbeidstid for menn er 37 timer, gjennomsnittlig boligareal for nye boliger er 129 kvadratmeter, gjennomsnittlig alkoholkonsum per voksen person per år er 6,2 liter ren alkohol, og så videre.

Gjennomsnittet er lett å beregne: Hvis jeg en fredag drikker tre glass vin, mens jeg lørdag drikker fire og søndag bare to, har jeg denne helgen i gjennomsnitt drukket tre glass per dag:

$$(3 + 4 + 2)/3 = 9/3 = 3$$
(aritmetisk gjennomsnitt)

De fleste vil si at dette gjennomsnittet gir en relativt god beskrivelse av mine drikkevaner denne helgen. Men hvis min kone drikker bare to glass på fredag, sju på lørdag og ikke noe på søndag, blir også hennes gjennomsnitt tre glass per dag. Allerede her ser vi at gjennomsnittet kan dekke over eller skjule store variasjoner:

Gjennomsnittet er kanskje ikke like egnet til å beskrive hennes drikkemønster denne helgen: De to like gjennomsnittstallene dekker over svært ulike fordelinger.

… men gir ikke alltid en god virkelighetsbeskrivelse

For å beregne et gjennomsnitt behøver man altså *ikke* å kjenne enkeltverdiene (hele fordelingen av en variabel), det er tilstrekkelig å kjenne totalen. Det er nok å vite at jeg i løpet tre dager har drukket ni glass vin: 9 dividert på 3 = 3. Gjennomsnittet er altså lett å beregne, men det gir ikke alltid en like god beskrivelse av virkeligheten. Gjennomsnittet sammenfatter og komprimerer et datamateriale (uten gjennomsnittet ville vi drukne i et hav av tall), men dermed mister vi også viktig informasjon om variasjonen eller spredningen.

La oss anta at vi har følgende data om månedslønnen til seks ektepar, som vist i tabell 1. Forskjellen i gjennomsnittsinntekt blir altså 5 400 kroner. Dette til tross for at den gjennomsnittlige forskjellen for de fem parene som tjener minst, bare er i underkant av 2 000 kroner. Fordi én mann har langt høyere inntekt enn de øvrige, blir altså forskjellen mer enn doblet. Gjennomsnittet har altså mange svakheter: Det er svært følsomt overfor ekstreme verdier, samtidig som det skjuler ulike fordelinger.

Gjennomsnittsberegninger gir også resultater som ikke finnes i virkeligheten; for eksempel har ingen familie 2,3 barn. I Statistisk sentralbyrå får vi ofte spørsmål om «gjennomsnittsnordmannen». En kollega pleier da å svare at han gjerne kan konstruere en slik, men at en slik person ikke finnes.

Tabell 1. **Gjennomsnittlig månedslønn**

	1	2	3	4	5	6	Gjennomsnitt
Menn	27 600	28 300	29 500	29 800	30 700	52 000	32 983
Kvinner	25 500	27 200	27 500	27 900	28 100	29 300	27 583
Differanse	2 100	1 100	2 000	1 900	2 600	22 700	5 400

Jan Erik Kristiansen er sosiolog og seniorrådgiver i Statistisk sentralbyrå, Formidlingsavdelingen. Han har lang erfaring i å presentere statistikk på en brukervennlig måte og har holdt en rekke kurs i statistikkforståelse og bruk av statistikk.

Han utga i 2007 boken «Tall kan temmes!» (IJ-forlaget), som delvis danner grunnlaget for denne spalten.

(jan.erik.kristiansen@ssb.no)

Alder: et spesialtilfelle

Hvis fem personer er henholdsvis 32, 45, 47, 51 og 62 år, skulle gjennomsnittsalderen bli:

$$(32 + 45 + 47 + 51 + 62)/5 = 47{,}4;$$
dvs. 47 år

Mens barn ofte oppgir alderen svært nøyaktig og med en tendens til å «strekke seg oppover»: «Jeg er sju og et halvt år» eller «Jeg er åtte, snart ni», ser vi som voksne litt større på det og runder av nedover. Det samme skjer i statistikken; en persons alder regnes oftest i antall fylte år. Men de fem personene vil sannsynligvis alle være noe eldre enn akkurat det oppgitte antallet fylte år (se tabell 2).

De fem er altså i gjennomsnitt 6 måneder eldre enn den oppgitte alderen. Derfor må vi legge 0,5 år til gjennomsnittsberegningen. Gjennomsnittsalderen for disse fem personene blir dermed 47,4 + 0,5 = 47,9 år; det vil si 48 år. Dette er et generelt prinsipp ved beregning av gjennomsnittsalder: Vi legger til 0,5 år, idet vi antar at fødselsdagene fordeler seg jevnt over året.

Tabell 2. **Beregning av gjennomsnittsalder**

A	32 år	4 måneder
B	45 år	7 måneder
C	47 år	3 måneder
D	51 år	10 måneder
E	62 år	6 måneder
Sum	**237 år**	**30 måneder**
Gjennomsnitt	47,4 år	6 måneder

Veid gjennomsnitt

Det aritmetiske gjennomsnittet behandler altså alle verdier likt. Hvis Per har en timelønn på 240 kroner timen og Kari 200, tjener de gjennomsnittlig 220 kroner. Men hvis mennene i en bedrift tjener 240 kroner timen og kvinnene 200 kroner, blir ikke gjennomsnittet nødvendigvis 220.

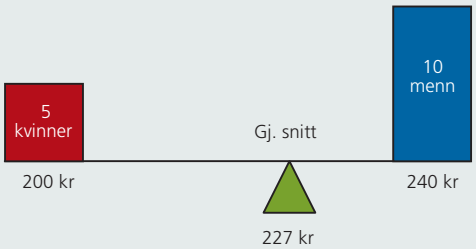
Når vi beregner gjennomsnitt for flere grupper, må vi ta hensyn til gruppenes størrelse. Avhengig av om det er flest menn eller kvinner blant de ansatte, vil gjennomsnittet ligge over eller under dette beløpet. Hvis det er 10 menn og 5 kvinner, blir gjennomsnittslønnen:

$$(240/15 \times 10) + (200/15 \times 5)$$
$$= 160 + 67 = 227$$

eller:

$$(240 \times 10) + (200 \times 5)/15 = 227$$

Dette veide gjennomsnittet kan også illustreres slik:



Fordi mennene er dobbelt så mange som kvinnene, får de størst vekt, og balansepunktet for vippen blir liggende nærmere mennenes timelønn enn kvinnenes.

Gjennomsnitt av hva?

Ofte vil man ved beregning av et gjennomsnitt ta utgangspunkt i hele populasjonen og også inkludere enheter i fordelingen som har verdien 0 på en variabel. For eksempel inngår også de som ikke har noen inntekt, i gjennomsnittsinntekten. Slik er det også når vi beregner gjennomsnittlig antall ferieturer per år eller gjennomsnittlig tid brukt til avislesing eller fjernsynsseing per dag. Men fordi det er en stor andel av befolkningen som ikke har vært på ferie eller lest aviser / sett på TV, ønsker vi også å se på gjennomsnittet bare for dem som har utført aktiviteten. Et eksempel:

I 2007 var det gjennomsnittlige antallet ferieturer for hele den voksne befolkningen 1,7. Men hele 21 prosent av befolkningen hadde ikke vært på noen ferietur dette året, noe som innebærer at de som hadde vært på ferie dette året, i gjennomsnitt hadde vært på flere turer, nemlig 2,1. Disse gjennomsnittene beregnes

som vanlig ved at man summerer antall turer og dividerer på antall personer i gruppen. Men hvis vi for eksempel vet at gjennomsnittlig antall turer for alle er 1,7, og at 79 prosent har vært på ferie, kan vi også beregne antallet turer for de ferierende slik:

$$1{,}7/79 * 100 = 2{,}1$$

og omvendt:

$$2{,}1/100 * 79 = 1{,}7$$

Barn og bydeler

I en liten by er det tre bydeler: A, B og C. Barnehagedekningen i de tre bydelene er henholdsvis 55,1, 72,2 og 86,5 prosent. I valgkampen sier så byens ordfører at den gjennomsnittlige barnehagedekningen i byen er 71 prosent. Dette tallet har han fått ved å beregne et gjennomsnitt for de tre bydelene, slik:

$$(55{,}1 + 72{,}2 + 86{,}5)/3 = 71{,}2$$

Men la oss nå si at fordelingen av antall barn og barnehageplasser i bydelene er som i tabellen nedenfor:

Tabell 3. **Barn og bydeler**

	Antall barn 1-5 år	Antall barn i barnehage	Barnehage-dekning
Bydel A	465	256	55,1
Bydel B	230	166	72,2
Bydel C	126	109	86,5
Hele byen	821	531	64,7

Bydel A, som har flest barn, har den laveste dekningsgraden, mens den minste bydelen (C), har høyest dekning. Når vi beregner et gjennomsnitt for bydelene, behandler vi alle bydelene likt (uveid gjennomsnitt), og de to små bydelene vil få like stor vekt som den store, og gjennomsnittet blir kunstig høyt. Om vi derimot beregner gjennomsnittet for hele byen på grunnlag av data om barna, får vi en lavere barnehagedekning, nemlig 64,7 prosent.

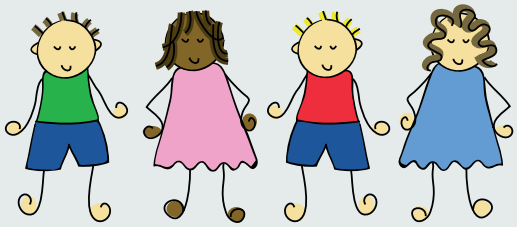
$$531/821 * 100 = 64{,}7$$

64,7 prosent kan også betraktes som et veid gjennomsnitt av prosentene for de tre bydelene, og kan beregnes slik:

$$(55{,}1/821 * 465) + (72{,}2/821 * 230) + (86{,}5/821 * 126)$$
$$= 31{,}2 + 20{,}2 + 13{,}3 = 64{,}7$$

Når ordføreren sier at gjennomsnittet for bydelene er 71 prosent, har han for så vidt sine ord i behold, men byens samlede barnehagedekning er altså lavere.

Mer generelt: Gjennomsnitt basert på aggregerte enheter som bydeler, kommuner eller fylker må ikke tas som uttrykk for gjennomsnittet på individnivå.



Gjennomsnitt av grupperte tall

For å beregne for eksempel gjennomsnittshøyden i en gruppe må vi altså enten kjenne verdiene (høyden) for hver enhet (person), eller vi må kjenne gruppens samlede høyde og antallet personer.

Men noen ganger foreligger slike opplysninger bare som en gruppert fordeling, som vist i tabell 4. Her kan vi beregne et (tilnærmet) gjennomsnitt på følgende måte: Vi antar at midtpunktet i hvert intervall representerer gjennomsnittshøyden i gruppen. For gruppen 165–169 setter vi midtpunktet til 167, i neste gruppe til 172, og så videre. De åpne intervallene «under 165» og «195 +» er mer problematiske, siden de ikke har noen nedre og øvre grense. Her må midtpunktet (gjennomsnittet) anslås – til 163 og 197 cm. Siden de åpne intervallene her omfatter svært få personer, spiller det relativt liten rolle hvor presise disse anslagene er.

Tabell 4. **Vernepliktige og høyde. 2005**

Høyde/cm	Fordeling	Midtpunkt	Sum
Under 165	1,2	163	195,6
165-169	4,8	167	801,6
170-174	15,0	172	2 580,0
175-179	26,9	177	4 761,3
180-184	27,9	182	5 077,8
185-189	17,0	187	3 179,0
190-194	5,8	192	1 113,6
195 +	1,4	197	275,8
	100		17 984,7

Kilde: Vernepliktsverket.

For hvert intervall multipliserer vi nå det antatte gjennomsnittet med antall personer eller den tilsvarende prosentandelen (om man her bruker absolutte eller relative tall, spiller ingen rolle). Deretter summerer vi den samlede høyden og dividerer med antallet personer. Resultatet blir en gjennomsnittshøyde på 179,8 cm.

Medianen

Et mål som ikke – på samme måte som gjennomsnittet – tar hensyn til ekstreme verdier, er medianen. Medianen er den verdien som deler en fordeling i to like store deler, slik at halvparten av enhetene ligger over og halvparten under denne verdien.

Dette målet brukes ofte når vi skal beskrive variabler hvor fordelingen ikke er symmetrisk, men skjev, som lønn, inntekt og formue. Men medianen brukes nok langt mindre enn det aritmetiske gjennomsnittet, blant annet fordi dette målet ikke er så intuitivt forståelig.

Heltidsansatte menns gjennomsnittslønn var i 2008 37 100 kroner per måned, sammenliknet med 32 300 for kvinner. Siden menns lønnsfordeling er langt skjevere enn kvinners (noen menn har svært høy lønn), kan medianen her være et vel så godt mål som gjennomsnittet. Medianen var 32 800 kroner for menn og 30 300 for kvinner. Målt på denne måten (når vi ser bort fra de ekstreme lønningene) øker kvinners lønn som andel av menns fra 87 til 92 prosent.

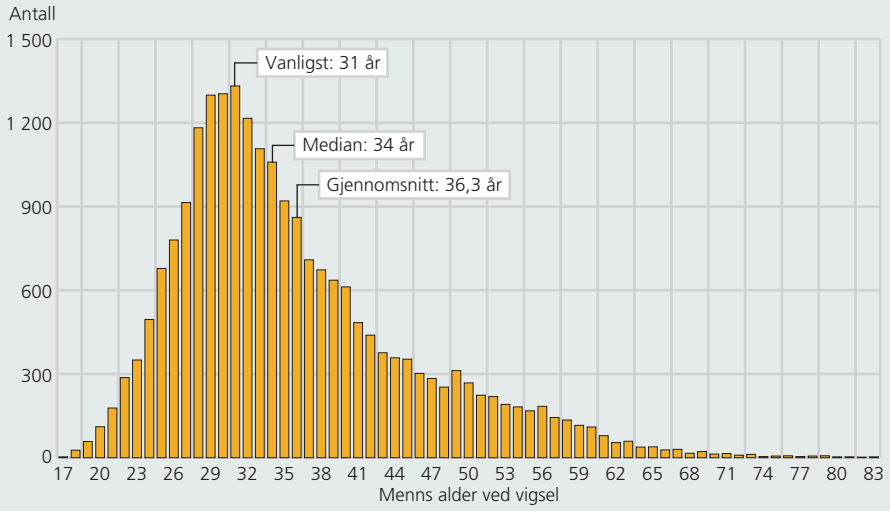
Typetallet

Et tredje mål for det vi kaller sentraltendensen i et datamateriale, er typetallet. Dette er den typiske eller vanligste verdien i en fordeling; det vil si den verdien som oftest forekommer. Dette målet er ikke så ofte brukt, men egner seg for enkelte variabler, som antall barn eller typisk ekteskapsalder.

Ulike mål gir ulike resultater

Figur 1 viser fordelingen av menns alder ved ekteskapsinngåelsen i 2004. Vi ser her hvordan de tre ulike «gjennomsnitte-

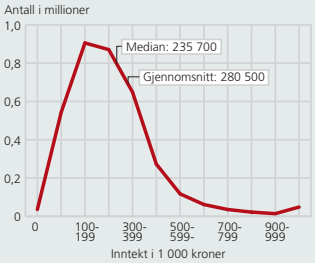
Figur 1. **Menns alder ved ekteskapsinngåelsen. 2004**



Kilde: Befolkningsstatistikk, Statistisk sentralbyrå.

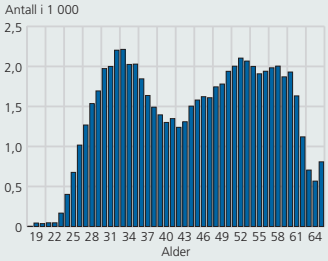
ne» gir forskjellige resultater: Den typiske eller vanligste ekteskapsalderen er 31 år. Gjennomsnittsalderen, derimot, er langt høyere – 36,3 år. Medianalderen, som ikke på samme måte som gjennomsnittet er påvirket av de mange giftelystne eldre, er 34 år. Rekkefølgen typetall < median < gjennomsnitt indikerer at fordelingen er høyreskjev, noe som også er typisk for mange andre fordelinger, som lønn, inntekt og formue.

Figur 2. **Inntektsfordeling for personer. Antall i 1 000. 2004**



Kilde: Selvangivelsesstatistikk, Statistisk sentralbyrå.

Figur 3. **Aldersfordeling for lærere i grunnskolen. 2004**



Kilde: Utdanningsspeilet 2005, Utdanningsdirektoratet.

Som en dromedars pukkel ...

Ser vi på inntektsfordelingen for personer, som vist i figur 2, er forholdet mellom gjennomsnitt og median omtrent som ovenfor: Gjennomsnittet er om lag 45 000 kroner høyere enn medianen. Og relativt sett er forskjellen større her, noe som henger sammen med at et stort antall personer tjener over 1 million kroner. Noen tilsvarende stor andel av eldre som inngår ekteskap etter for eksempel fylte 80 år, finnes ikke, siden det her finnes en naturlig øvre grense. En slik grense finnes ikke når det gjelder inntekt.

Selv om begge de to fordelingene ovenfor er skjeve, ligger likevel tyngdepunktet (det vil si flertallet av observasjonene) mer eller mindre rundt midten av fordelingen, som likner en dromedars pukkel.

... eller som en kamels

Hvis vi ser på aldersfordelingen for lærere i grunnskolen, som vist i figur 3, er gjennomsnittlig alder 45 år. Her ser vi et eksempel på en litt annen type fordeling, som likner mer på puklene til en kamel. Men også i dette tilfellet må gjennomsnittet sies å være lite typisk, idet de to tyngdepunktene i fordelingen ligger godt under og over dette. Heller ikke medianen (46 år) gir noe bedre signalement på en typisk lærer, som mest sannsynlig er i begynnelsen av 30-årene eller i begynnelsen av 50-årene.

Hvilket mål skal vi så bruke?

En mulighet er selvfølgelig å presentere alle tre målene. Men dette blir fort mange tall, særlig hvis man skal sammenlikne to eller flere grupper (for eksempel ekteskapsalder eller lønn for menn og kvinner). Svaret må bli at bruk av gjennomsnitt eller median må vurderes i hvert enkelt tilfelle, ut fra det en vet om fordelingen. I tilfeller hvor fordelingen er veldig spiss, og én enkelt verdi peker seg ut som den vanligste, bør også typetallet oppgis.