### Exempel på problem med vanliga deskriptiva informationsmått

De två vanligaste måtten på vilka värden någon kvantitet får i ett stickprov, eller helt enkelt bland en mängd observationer, är medelvärdet och variansen (och dess kvadratrot, som kallas standardavvikelse). Medelvärdet (eller mer exakt, det aritmetiska medelvärdet) ger för de flesta den intuitionen att man betraktar det typiska, eller genomsnittliga värdet hos individerna i datamaterialet. Låt $x_i$ vara det observerade värdet på den intressanta kvantiteten hos en individ $i$ i ett datamaterial som består av totalt $n$ individer. Då beräknas medelvärdet ($\bar{x}$) enligt följande formel:

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

**Ett exempel**. Vi hör följande påstående: "Finländare dricker i genomsnitt tre-fyra öl om dagen". Detta påstående är baserat på följande uppgifter från ett datamaterial, där $X$ betecknar antalet öl man dricker per dag.

| $X$ | % |
|---|---|
| 0 | 40 |
| 1 | 10 |
| 2 | 0 |
| 3 | 0 |
| 4 | 0 |
| 5 | 1 |
| 6 | 5 |
| 7 | 40 |
| 8 | 4 |
| 9 | 0 |

I tabellen ovan står "%" för procentandelen individer i datamaterialet som dricker ett visst antal öl. Medelvärdet för detta material är lika med 3.57, vilket motsvarar påståendet om "sådär tre-fyra om dan". Hur bra beskriver då påståendet det vi ser i data? Rätt dåligt, eftersom ingen verkar dricka "tre-fyra om dan". Man måste vara försiktig då man pratar om "det genomsnittliga", eftersom medelvärdet är en meningsfull beskrivning om fördelningen av värden om de i verklighet ligger rätt symmetriskt runt det. I annat fall (som i exemplet ovan) kan medelvärdet vara ett vilseledande mått på kvantitetens fördelning. Detta gäller alltid om fördelningen har flera "toppar" än bara en. I statistiska termer säger man att medelvärdet bygger på

antagandet om en unimodal och symmetrisk fördelning som representarar de observerade värden i datamaterialet. Medelvärdet hade varit okej, om tabellen sett t. ex. ut såhär:

| $X$ | % |
|---|---|
| 0 | 4 |
| 1 | 5 |
| 2 | 8 |
| 3 | 15 |
| 4 | 40 |
| 5 | 15 |
| 6 | 7 |
| 7 | 4 |
| 8 | 2 |
| 9 | 0 |

Utöver medelvärdet brukar man använda varians (eller standardavvikelse) som ett spridningsmått för att beskriva en kvantitets fördelning. Variansen ($s^2$) kan räknas ut såhär:

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

dvs. standardavvikelsen är då

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

Variansen beskriver alltså spridningen av värden runt medelvärdet (den representerar det genomsnittliga kvadrerade avståndet från medelvärdet). Precis som medelvärdet, är även variansen en dålig beskrivning om fördelningen är kraftigt asymmetrisk eller har flera "toppar". Summan av kardemumman är att medelvärdet och variansen fungerar bra om datamaterialet är fördelat så att det ungefär liknar en normalfördelning. För den första tabellen skulle variansen vara 11.57. Detta representerar en grov överskattning av spridningen i datamaterialet, eftersom fördelningen i själva verket är kraftigt koncentrerad på värden 0 och 7. En acceptabel beskrivning av fördelningen i ord kunde istället vara t. ex. "ca hälften av finländare dricker inte alls öl dagligen och den andra halvan av befolkningen dricker ungefär sju öl om dagen".

**Dagens budskap**:

Använd alltid visuella medel (= statistisk grafik) utöver enkla deskriptiva numeriska mått för att beskriva dina datamaterial! Visuella beskrivningar är oslagbara och man kan dessutom då lättare bedöma om medelvärden, varianserna osv. är pålitliga.