Föreläsning 30/8

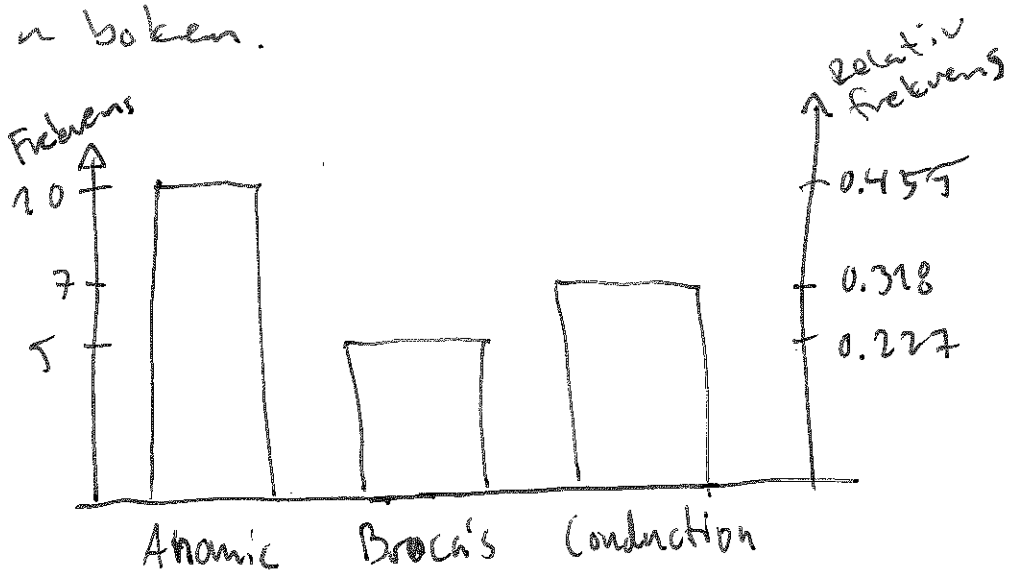
(kapitel 2.1) Beskrivning av kvalitativ (kategorisk) data

Först några ord. En klass är en av kategorierna av kvalitativa data kan uppdelas i. Klassfrekvensen är antalet observationer från ett dataset ~~eller stickprov~~ som tillhör en viss klass. Den relativa klassfrekvensen är klassfrekvensen delat på ~~det~~ det totala antalet observationer.

brukar betecknas med n

Klassprocent = 100 x Relativa klassfrekvensen.

Vi pratade om cirkeldiagram förra föreläsningen.
Man kan även använda sig av stapeldiagram.
Man kan då antingen använde sig av klassfrekvens eller relativ klassfrekvens. Exempel; Figur 2 från boken.

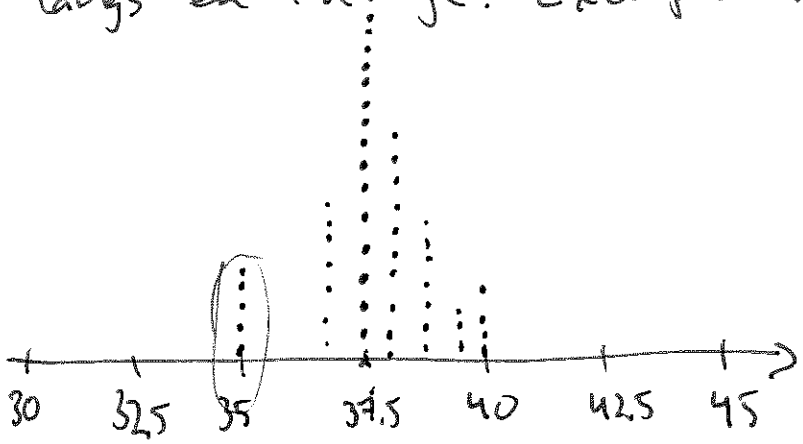


Paretodiagram nämns även i boken.

Beskrivning av kvantitativ (numerisk) data (kapitel 2.2)

I boken beskrivs tre olika sätt att grafiskt presentera kvantitativ data: dot-plot, stem-and-leaf-displays och histogram.

I en dot-plot så grupperar man datapunkter som ligger nära varandra och "staplar" dem på varandra längs en tallinje. Exempel från boken: Figur 8.



Betyder att det är 5 observationer av 35

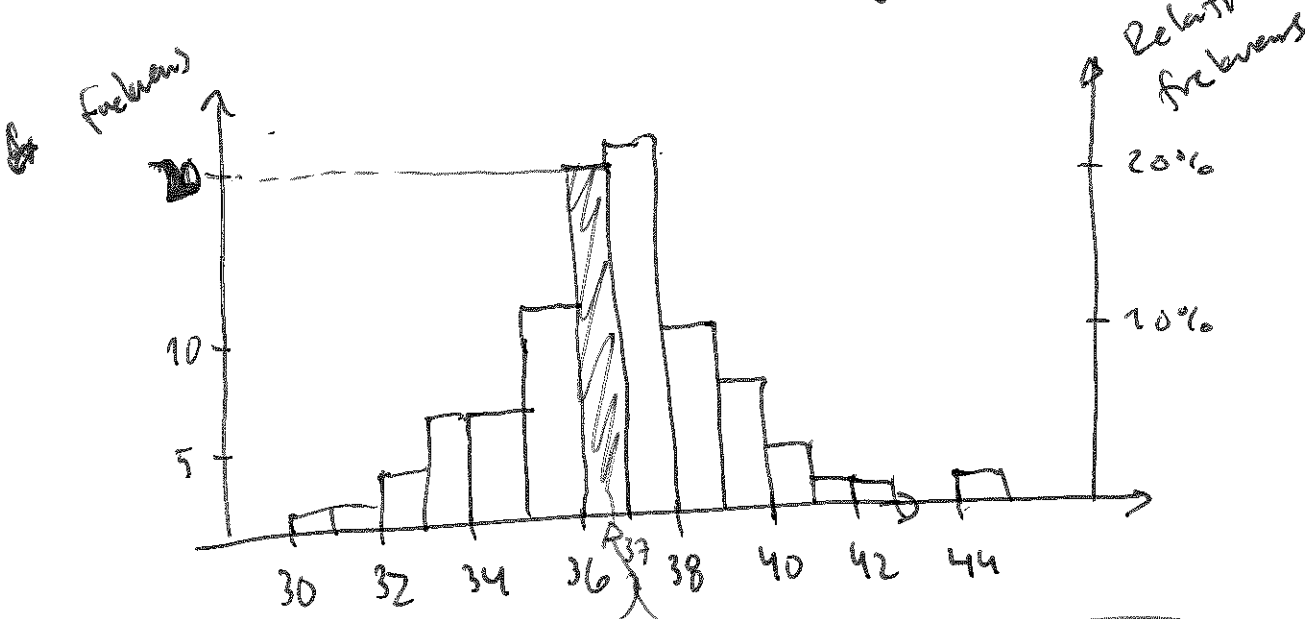
Stem-and-leaf-diagram staplas åt andra hållet, se figur 9 i boken:

30	0
31	8
32	5799
33	126899
⋮	

Detta betyder att vi har 4 observationer mellan 32 och 33, nämligen 32.5, 32.7, 32.9 och 32.9.

Första kolumnen motsvarar siffrorna innan decimaltecknet, andra kolumnen motsvarar siffrorna efter decimaltecknet.

Det mest använd av de två sätten är <u>histogram</u>.
När man gör ett histogram så delar man in datan i intervall. Exempel från boken: Figur 10.



Detta innebär att det finns 20 observationer mellan 36 och 37

Utseendet på ett histogram beror på hur fin klassindelningen är, dvs hur många små intervall man använder.

Summa - notation (kapitel 2.3)

Det är väldigt vanligt att använda bokstäver för att representera tal + ex $x_1, x_2, \ldots, x_n$. För att beskriva summan av en samling tal brukar man använda sig av summanotationen:

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + x_3 + \ldots + x_n$$

Ex: Om $n=5$ och $x_1=5, x_2=3, x_3=8, x_4=5, x_5=4$ så har vi att

$$\sum_{i=1}^{5} x_i = 5+3+8+5+4 = 25$$

6 Man kan också summera funktioner av talen.

Ex: Samma x och x som i förra exemplet

$$\sum_{i=1}^{5} x_i^2 = 5^2 + 3^2 + 8^2 + 5^2 + 4^2 =$$

$$= 25 + 9 + 64 + 25 + 16 = 139$$

Lägesmått (kapitel 2.4)

Olika typer av lägesmått används för att beskriva ungefär hur stora värden som finns i ett dataset. Det finns i huvudsak tre olika lägesmått:

Medelvärdet av ett dataset är summan av mätvärdena delat med antalet mätvärden:

I boken gör dom skillnad på populationsmedelvärde (μ) och stickprovs medelvärde. ($\bar{x}$)

Medianen av ett dataset är det mittersta värdet när man sorterat datan.

Typvärdet av ett dataset är det mest förekommande värdet.

Observera att dessa tre värden inte alltid är lika. I datasetet från kapitel 2.2 så är medelvärdet 36.944 medan medianen och typvärdet är 37.

Spridningsmått (kapitel 2.5 + 2.6)

Det är även väldigt vanligt att presentera mått på hur mycket variation det finns i datan, de används spridningsmått. De vanligaste spridningsmåtten är Stickprovsvariansen och stickprovsstandardavvikelsen.

Stickprovsvariansen:

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

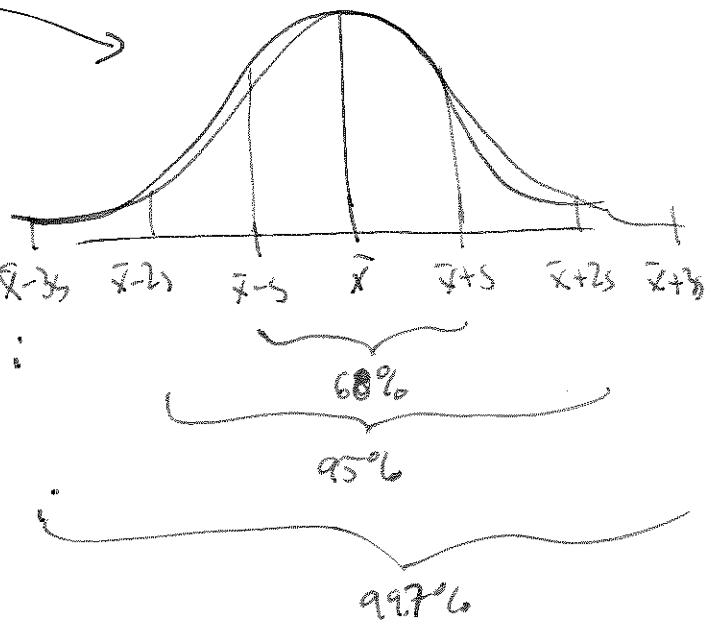
OBS: För att ta fram s^2 måste vi först ta fram $\bar{x}$

Stickprovsstandardavvikelse: $s = \sqrt{s^2}$

~~Fördelen med~~

En av fördelarna med standardavvikelsen över variansen är att den har samma enhet som datan. Om vi mäter längden i cm på en grupp människor så kommer s att vara den förväntade avvikelsen i cm från medelvärdet.

En annan fördel är följande: om datan ser ut ungefär som en klocka så har vi att



→ ca 68% av datan ligger i intervallet $(\bar{x}-s, \bar{x}+s)$

→ ca 95% av datan ligger i intervallet $(\bar{x}-2s, \bar{x}+2s)$

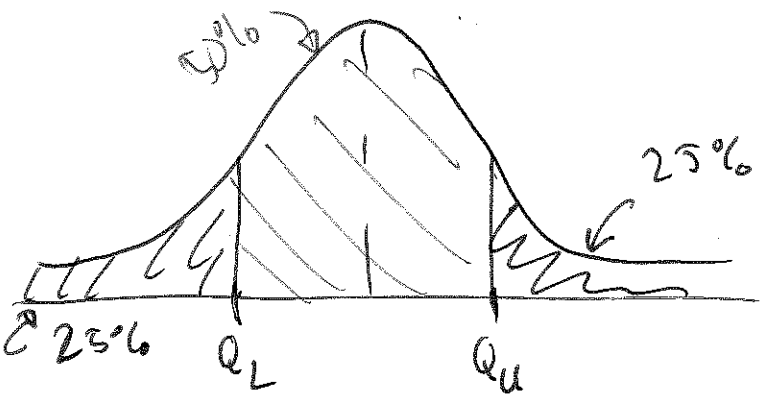
→ ca 99.7% av datan ligger i intervallet $(\bar{x}-3s, \bar{x}+3s)$

Percentiler ~~Kvartiler~~, outliers, box plots (kapitel 2.7+2.8)

För ett tal p mellan 0 och 100 ~~är~~ är den p:te percentilen det tal för vilket p% av mätvärden är lägre än det talet. Tex om vi har 200 mätvärden och 20% av mätvärdena är mindre än 10 så är den 20:de percentilen = 10.

Den <u>nedre kvartilen</u> Q_L är den 25 percentilen och den <u>övre kvartilen</u> Q_U är den 75 percentilen. Alltså är 50% av alla mätdata mellan Q_L och Q_U.

Kvartilavstånd:
$IQR = Q_U - Q_L$



<u>Outliers</u> är observationer som ligger extremt långt från resten av datasetet. Det kan finnas lite olika anledningar till att man har sådana i sitt dataset:

1. Mätfel
2. Observationen kommer från en annan population
3. Observationen är oväntad

Det kan vara väldigt svårt att avgöra orsaken!

En box-plot ser ut på följande sätt. (Figur 31)