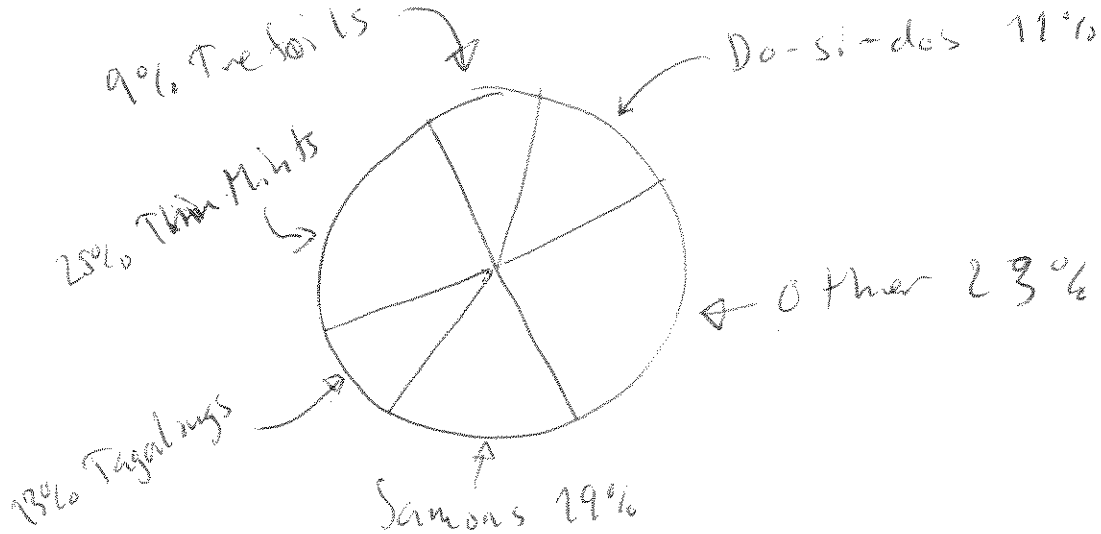
# Statistik (kap 1)

Man brukar dela in begreppet statistik i två: deskriptiv statistik och statistisk inferens.

Deskriptiv statistik använder numeriska och grafiska metoder för att sammanfatta informationen om data och att presentera denna information på ett bra sätt. Exempel från boken, cirkeldiagram: "Study 1" om flickscouters kakförsäljning.

Ca 150 miljoner kaklådor säljs varje år och de kan delas in i olika kategorier (smaker).



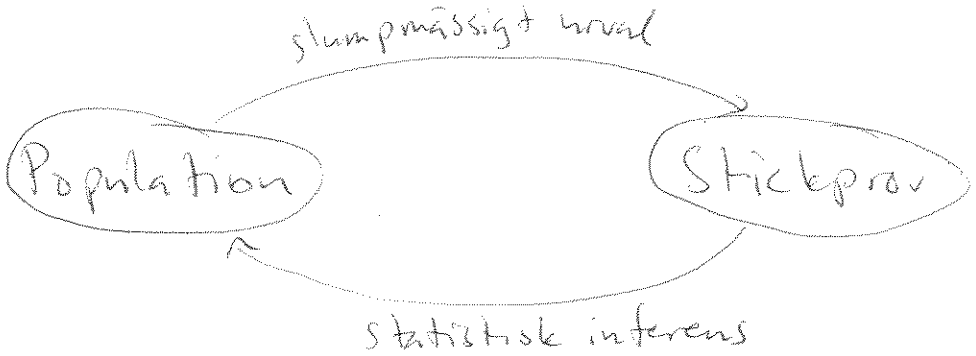

Statistisk inferens använder stickprov för att göra uppskattningar, ta beslut, göra förutsägelser och andra generaliseringar om ett större dataset. Exempel från boken: "Study 2" om TV- och datorspelandets effekter på personers visuella uppmärksamhets förmåga.

Läs "study 2"!

I statistiska undersökningar använder man sig ofta av både deskriptiv statistik och statistisk inferens. Man måste beskriva och presentera datan (deskriptiv) men man måste även kunna dra slutsatser (inferens).

Inom statistisk inferens vill man oftast dra slutsatser om en population med hjälp av ett stickprov från populationen. En population kan i detta ~~sammanhang~~ sammanhang t ex vara: alla skattebetalare i Sverige, alla rösträttsgrade invånare i Göteborg, alla bilar som produceras på volvo, Torslanda.

Ett stickprov ska alltid väljas ut slumpmässigt utifrån hela populationen, men stickprovet utgör aldrig hela populationen.

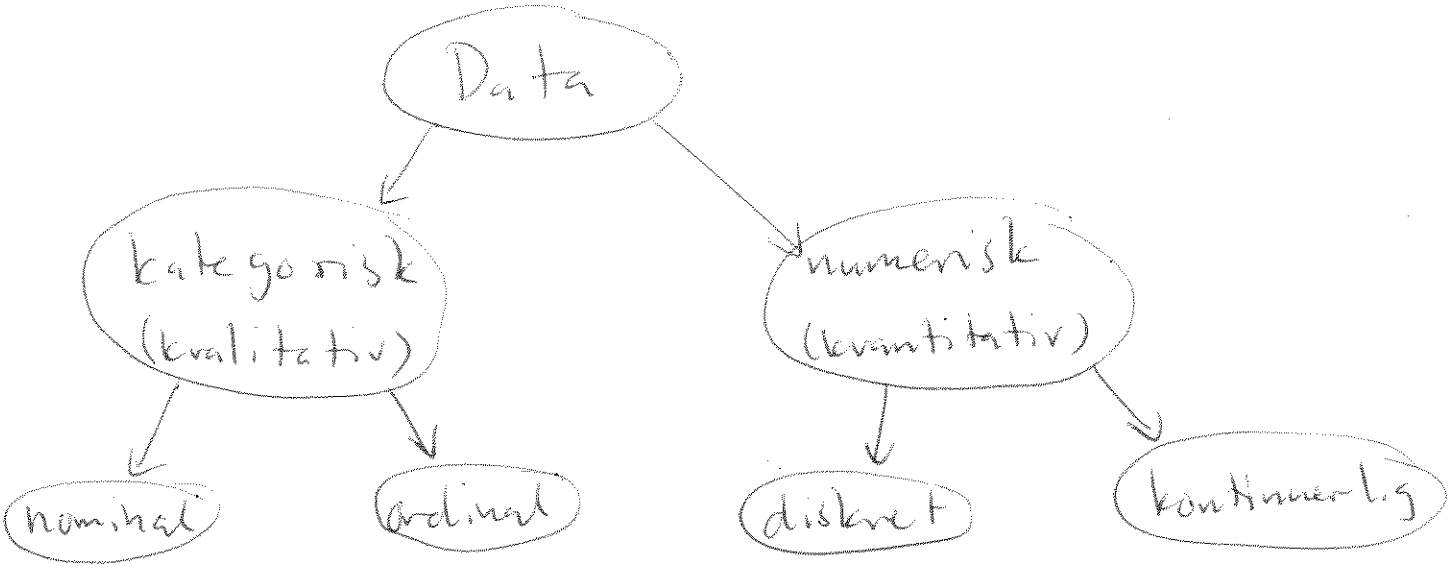


I boken presenteras fem element inom statistisk inferens:

1. Population
2. Variabler (Vad är det vi undersöker)
3. Stickprov
4. Inferens (Hur har vi gjort analysen? Vad är slutsatsen?)
5. Variationsmått (Hur säkra är vi på vår slutsats?)

---

## Typer av data



Exempel:

En mäklarbyrå vill sammanfatta deras sålda bostäder under ett år. Många olika datatyper att titta på

nominal: typ av bostad (fristående hus, radhus, lägenhet)

ordinal: skalor (enkätsvar, nöjdhetsindex)

diskret: antal (antal badrum, sovrum)

kontinuerlig ej uppräkneelig (boyta, kostnader)

Vilka statistiska metoder som används påverkas av vilken typ av data vi är intresserad av. Vi kan t ex inte beräkna ett medelvärde av nominal data, och vi kan inte göra ett cirkeldiagram av kontinuerlig data.

---

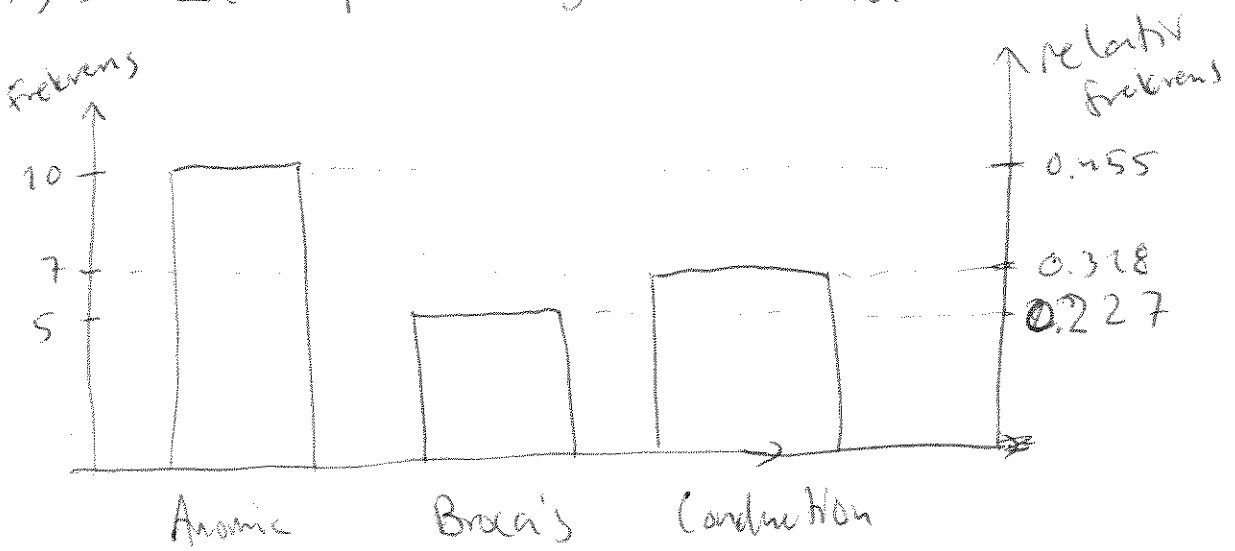
(kap 2.1)

## Beskrivning av kvalitativ (kategorisk) data

Först några ord. En klass är en av kategorierna som kvalitativa data kan delas upp i. Klassfrekvensen är antalet observationer från datasetet som tillhör en viss klass. Den relativa klassfrekvensen är klassfrekvensen det totala antalet observationer och klassprocent = 100 x relativa klassfrekvensen.

Förutom cirkeldiagram kan man även använda sig av stapeldiagram för att illustrera kvalitativ data.

Man använder sig då av antingen klassfrekvensen eller den relativa klassfrekvensen (eller klass-procent). Exempel: figur 2 från boken

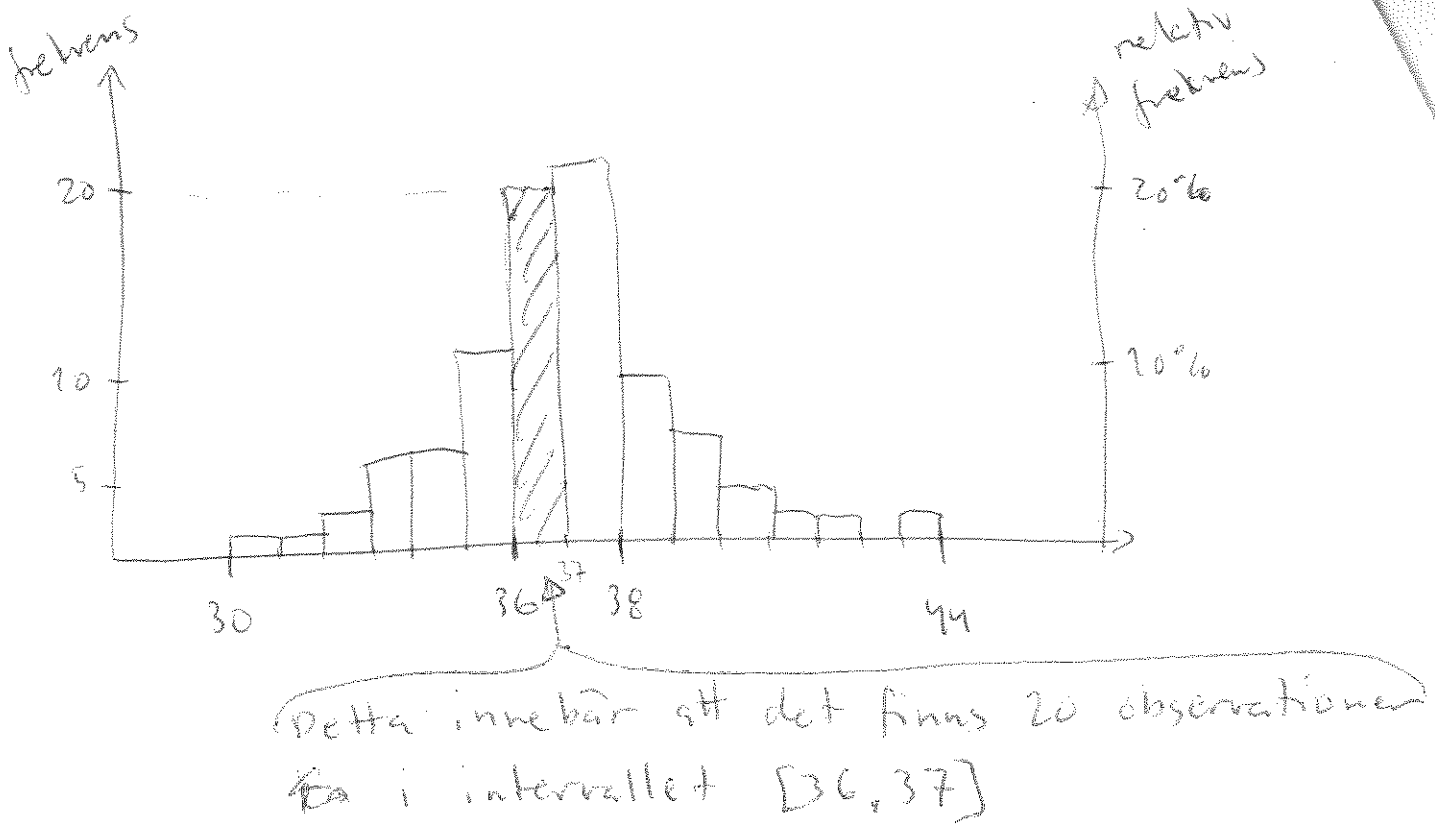



---

## Beskrivning av kvantitativ (numerisk) data (kap 2.2)

I boken beskrivs tre olika sätt att grafiskt presentera kvantitativ data: dot-plot, stem-and-leaf-displays och histogram. Det överlägset mest använda av dessa är histogram och det är det vi kommer gå igenom i denna kurs.

När man gör ett histogram så delar man in datan i intervall. Exempel från boken: figur 10.




Detta innebär att det finns 20 observationer i intervallet [36, 37]

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

---

## Lägesmått (kap 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 på antalet mätvärden $\bar{x} = \frac{1}{n}\sum_{k=1}^{n} x_k$

**Medianen** är det mittersta värdet när datan är sorterad

**Typvärdet** är det mest förekommande värdet.

(~~Bara~~ används bort för kontinuerlig data)

Observera att värdet på dessa inte är samma. I datasetet från kapitel 2.2 så är medelvärdet 36.994 medan medianen är 37.

## Spridningsmått (kap 2.5 - 2.7)

Spridningsmått används för att se hur mycket variation det finns i datan. Det finns tre viktiga spridningsmått:

Stickprovsvariansen: $s^2 = \frac{1}{n-1} \sum_{k=1}^{n} (x_k - \bar{x})^2$

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

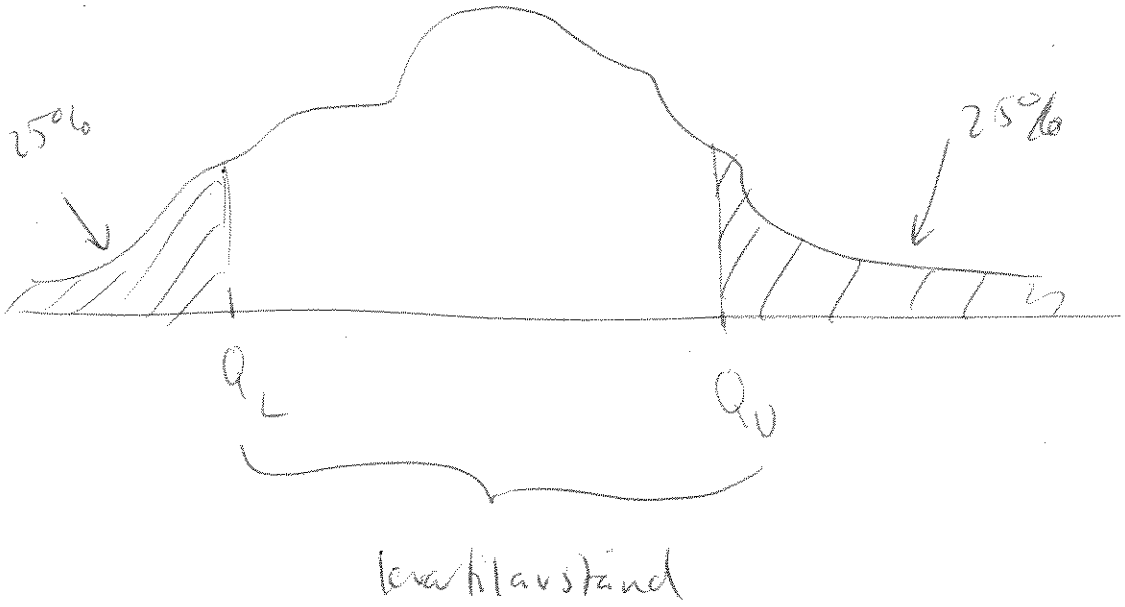
Kvartilavstånd (återkommer till detta snart)

Stickprovsvariansen och stickprovsstandardavvikelsen är uppenbarligen väldigt lika, och de är relaterade till varandra och standardavvikelsen ~~är~~ av en slumpvariabel (nästa vecka).

Den nedre kvartilen $Q_L$ är det värdet som är större än exakt 25% av datan, den övre kvartilen $Q_U$ är det värde som är större än exakt 75% av datan.

$$\text{Kvartilavstånd} = \underset{\text{"inter quartile range"}}{IQR} = Q_U - Q_L$$

OBS: 50% av datan ligger mellan $Q_L$ och $Q_U$


25%
25%
$Q_L$
$Q_U$
Leachlavstand