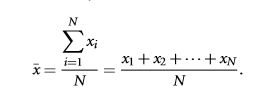
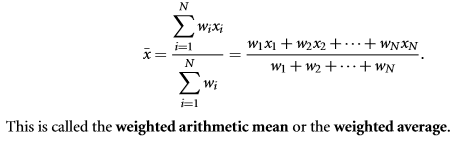
Univariate = samling af data der har en attribut, bivariate=to osv.

Central tendency:

1. Mean (Average value)
2. Median (middle value)
3. Mode (Most frequent value)
4. Midrange (Average of smallest and largest value)



1 Mean

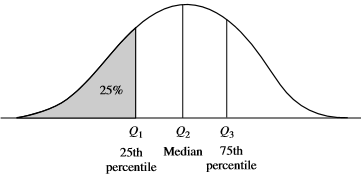


Attributes:

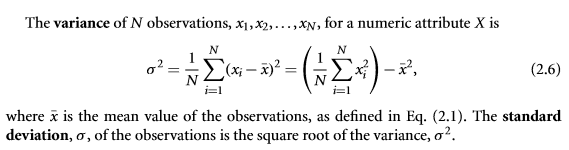
* Nominal: Tekst (ikke sammenligneligt matematisk), f.eks. job (programmør, babysitter osv.)
* Binary: Boolean, har to værdier, behøver ikke være true false, men er det ofte.
* Ordinal: Tekst: sammenlignelige ved brug af median og mean, men ikke Mode, da det ikke er en definerbar skala.
* Numeric: TAL: to typer:
  + Interval-scaled: sammenlignelige ved brug af median, mean og mode, men har ikke noget 0 punkt og en værdi kan derfor ikke siges at være ”dobbelt så stor” (f.eks. celsius skalaen)
  + Ratio-scaled: -.- Har en nul-punkt, som f.eks. kelvin skalaen.

Measuring describution:

* Range: max() – min().
* Quantiles: Gruppe af values i rækkefølge
* Quartiles: Quantiles opdelt i 4 grupper
* Percentile: quantiles opdelt i 100 grupper.
* Inter-quartile range: Q3-Q1 (de midterste to quartiles)



The **ﬁve-number summary** of a distribution consists of the median (Q2), the quartiles Q1 and Q3, and the smallest and largest individual observations, written in the order of Minimum, Q1, Median, Q3, Maximum



# Forelæsning

Hvorfor er “whiskers” på boxplot ikke lige lange?

Histogram bin laves vel fra numerical til ordinal.

## Preprocessing

Data cleaning.

Missing data:

* Ignore object (Might remove important tuples)
* Fill in manually (expensive)
* Global constant (Be vary of algorithms thinking there’s a pattern with constants)
* Tendency measure to “best guess”
  + Might introduce bias in the data set.

Data smoothing:

* Remove noise from the data. This is especially relevant in physical measurement.
* Binning values. AKA cluster values alike.
  + Types of bins are: bin mean bin medians, bin boundaries.
* Regression function to smooth data.

Outlier analysis:

* Do cluster analysis to identify the outliers.
* This can be achieved by ex. Scatterplotting the data.

Data redundancy:

* Elimenate redundant attribute, ex. If the value of an attribute cn be derived from some other attribute(s).
* Use visual means like scatterplot/correlation analysis.
* Nominat data => chi-square.

Data integration:

Data transformation:

