	L1	L2	L3	L4	Sum
young					100
medium					100
old	35,5	46.2	14.0	4.3	100

(LSU 3 Nov 2021)

35,5% of the oldest age group have education level 1.

Scales of measurement

discrete: Nominal, Ordinal

discrete or continuous: Metric (Interval, Ratio)

increasing information

quality problem

... broken: yes/no

... degree of problem: none - moderate - severe

... degradation metric measurement

Measures of central tendency

Nominal	Ordinal	Metric
Mode	Mode	Mode (sometimes difficult for continuous data)

problematic, if the distribution is very flat

A1	A2	A3
20	20	19

Median $x_{0,5}$ (Ordinal)
50%-quantile
middle value

Median $x_{0,5}$ (Metric)
50%-quantile

$$x_{0,5} = \begin{cases} x_{(\frac{n+1}{2})} & n \text{ odd} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2} & n \text{ even} \end{cases}$$

The median is robust

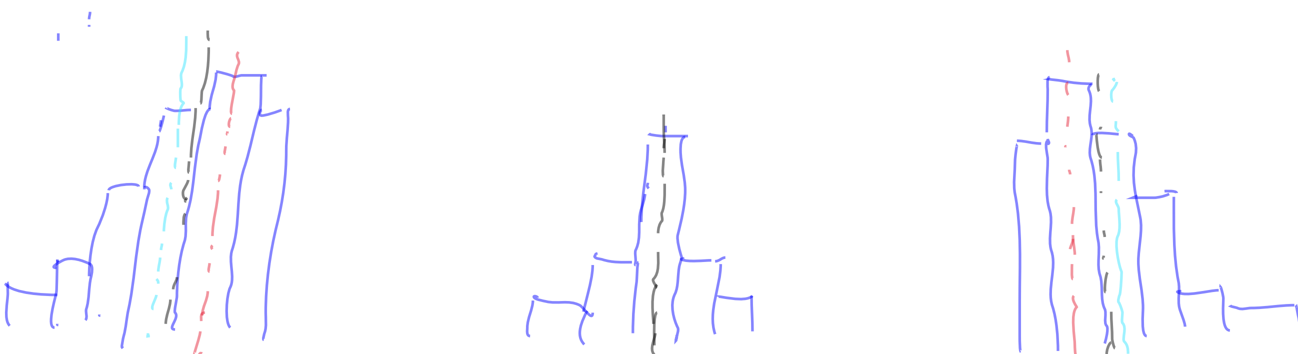
$x_{(1)}$ = smallest data value .... $x_{(n)}$ largest data value

Measures of central tendency only for metric data:

Arithmetic mean: $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ unrobust

Robustified version: trimmed mean
omit a percentage of values at both ends of the scale

Relation between mode, median and mean can be used for assessing shape of a distribution:



left-skewed
mode > median > mean

symmetric
mode = median = mean

right-skewed
mode < median < mean

Measures of variability (or "dispersion")

only for metric variables

Range: $x_{(n)} - x_{(1)}$ (largest − smallest value)

↖ remember: parentheses in index denote ordered values

very unrobust

For the next metric, we need to define quantiles: The p-quantile x_p is a value that separates the smallest p from the largest $1-p$
(p a proportion between 0 (0%) and 1 (100%))

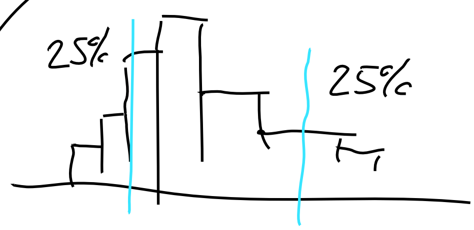
We already saw $x_{0.5} = x_{50\%}$, the median.

One (of many) formula for x_p:

$$x_p = \begin{cases} x_{(k)} & n \cdot p \text{ not an integer, } k = \lceil n \cdot p \rceil \ (\text{ceil}(n*p)) \\ \frac{x_{(k)} + x_{(k+1)}}{2} & n \cdot p = k \text{ an integer} \end{cases}$$

$$IQR = x_{0{,}75} - x_{0{,}25}$$

width of the middle 50% of the data



robust

upper and lower quartile

Mean absolute deviation from mean

$$\frac{1}{n} \sum_{i=1}^{n} |x_i - \bar{x}|$$

Mean absolute deviation from median

$$\frac{1}{n} \sum_{i=1}^{n} |x_i - x_{0{,}5}|$$

Median minimizes $\sum_{i=1}^{n} |x_i - m|$ w.r.t. m

Shortest average distance from a center value that is achievable with these data

Median absolute deviation from median

...

Empirical variance:

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

Here, $\bar{x}$ minimizes $\sum_{i=1}^{n}(x_i - m)^2$ w.r.t. m

This is one reason that $\bar{x}$ is so unrobust: square emphasizes large deviations

Variance

- is very important
- is very unrobust
- is in squared units of the data
- hard to interpret
- moment of inertia

Standard deviation

- square root of variance
  $\Rightarrow$ is again in units of the data