(Stat. Comp. 10 Nov 2021)

Visualization of a distribution of metric data

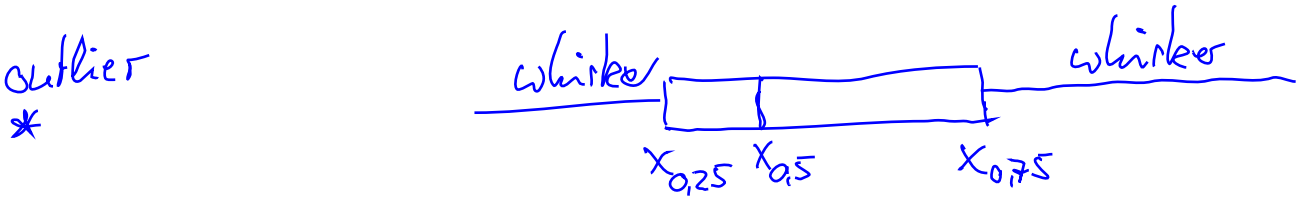
Boxplot: based on quantiles

Quantile: prop. p of data smaller | prop. $1-p$ of the data larger

x_p

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

Quartiles: $x_{0,25}$, $x_{0,5}$, $x_{0,75}$



Fences (DO NOT DRAW THEM):
- Width of box = IQR
- lower fence: $x_{0,25} - 1.5\,IQR$
  upper fence: $x_{0,75} + 1.5\,IQR$

define area in which we expect "normal" data

Whiskers:
lines from box to outmost data points within the fences

Outliers:
data points outside of fences

Example: Swiss Fertility

$n = 47$ $p = 0,25$

$n \cdot p = 11.75$

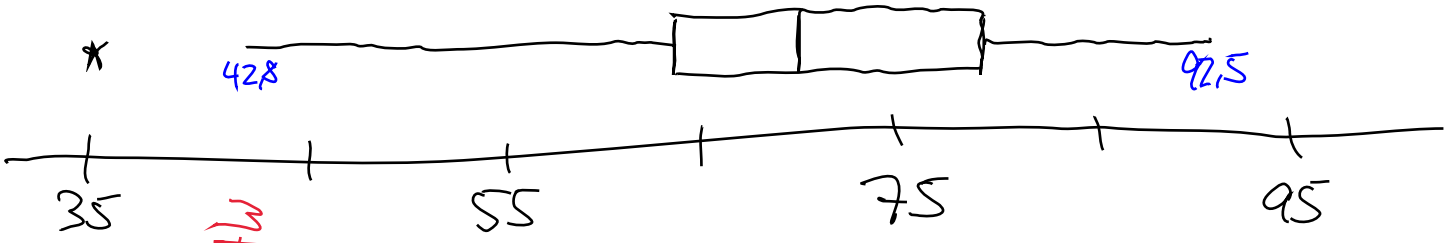
$x_{0.25} = x_{(12)} = 64.4$

$x_{0.5} = x_{(24)} = 70.4$

$x_{0.75} = x_{(36)} = 79.3$

$n \cdot p = 23.5$

$n \cdot p = 35.25$



DON'T DRAW

lower fence: $64.4 - 22.35 = 42.05$

$IQR = 79.3 - 64.4 = 14.9$

$1.5 \cdot IQR = 22.35$

upper fence: $79.3 + 22.35 = 101.65$

whiskers: left: 42.8

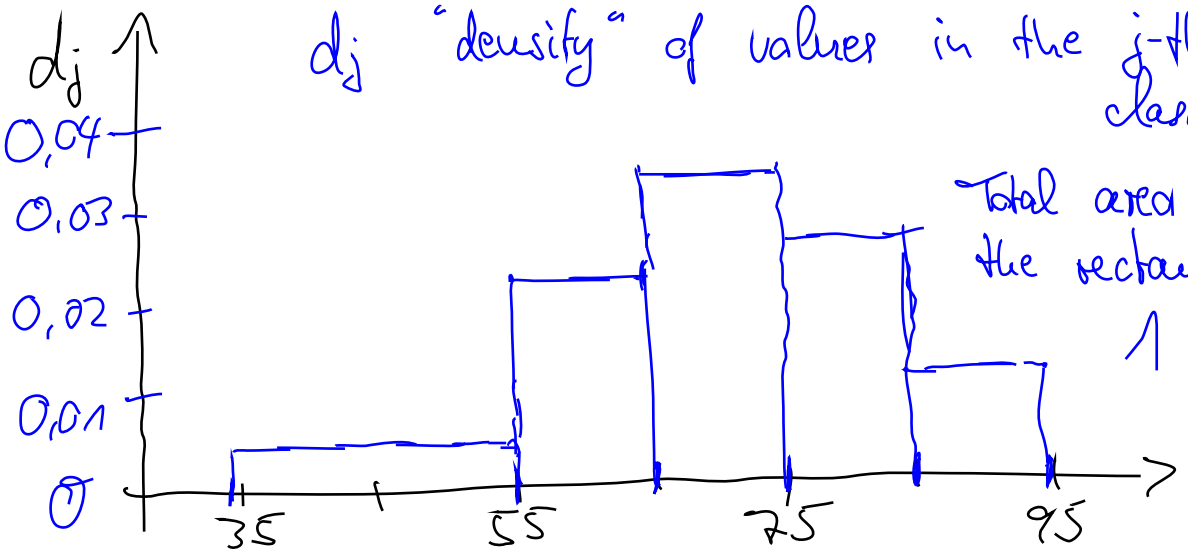
right: 92.5

Histogram and kernel density for continuous quantitative variables

Class the data (again fertility):

Class j	$> x_j^l$	$\leq x_j^u$	width$_j$	n_j	$h_j = n_j/n$	$d_j = h_j / \text{width}_j$
1	34	55	21	4	0.085	≈ 0.004
2	55	65	10	10	0.213	0.0213
3	65	75	10	16	0.340	0.034
4	75	85	10	12	0.255	0.0255
5	85	95	10	5	0.106	0.0106
				47	~ 1.000	

Histogram presents relative frequencies by rectangle areas.



d_j "density" of values in the j-th class

Total area of the rectangles: 1

d_j = height need for class j so that area of rectangle becomes h_j

Area of a rectangle:

A = height × width

$\Rightarrow$ height = A/width

R: function hist

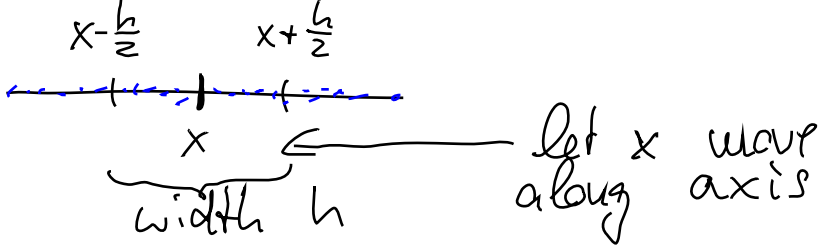
argument freq = FALSE ensures area 1

Disadvantage of histogram: not continuous, jumps

Kernel density estimate solves this problem:

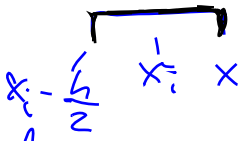
kde

- histogram uses several fixed non-overlapping "windows", the classes
- kde has a moving window for each x on the horizontal axis:



Simplest: count the values in the window ⇒ relative frequency
↳ normalize it so that the result has area 1 under the curve

Think of this approach in terms of a rectangular kernel:

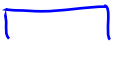
Place a rectangle 

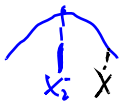
around each and every data value x_i, $i = 1, \dots, n$

For a value x on the axis, add up all the rectangle values from x_i's for which $x \in [x_i - \frac{h}{2}, x_i + \frac{h}{2}]$.

It's the same as above.

Generalizes to more useful kernels


rectangular = unsmooth



↳ smoother result esp. for smaller data sets