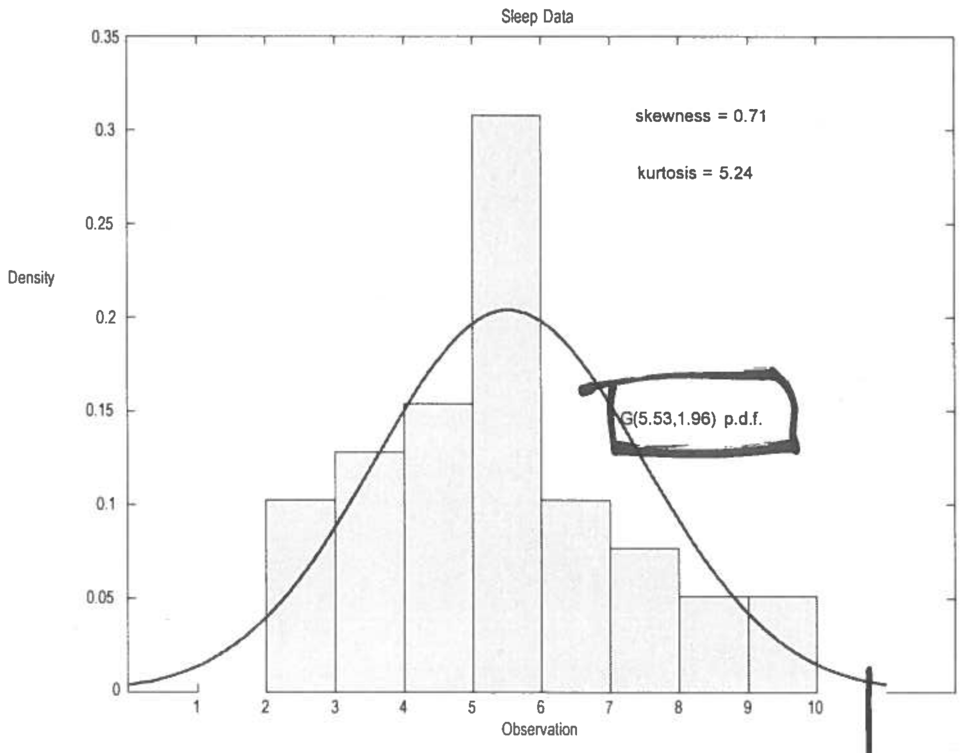
STAT 231

Roadmap

- 5 min recap of last class
- Graphical Data Summaries
    - Density Histogram ✓
    - Box-PLOT
    - Empirical cdf
    - Scatter plot

- find the properties of the data set
- "identify" the distribution from which the data is drawn.



Skewness = 0·71

Kurtosis = 5·24

G(5·53, 1·96)

mean s.d.

Best possible Normal (Gaussian) distribution that "fits" the data set.

Q1 Is the data right skewed?

- (i) Right - Skewed.
- (ii) Left Skewed
- (iii) Symmetric.

Q2 : Is the normal approximatio appropriate?

- (i) Yes
- (ii) No

Q3 : Find which group the median is?

Left as Exercise.

- **Empirical cdf**.

Data: $\{y_1, \ldots y_n\}$

$y_1 \leq y_2 \ldots \leq y_n$

Cumulative Distⁿ function

↑

Definition : $F(y) =$ Empirical cdf

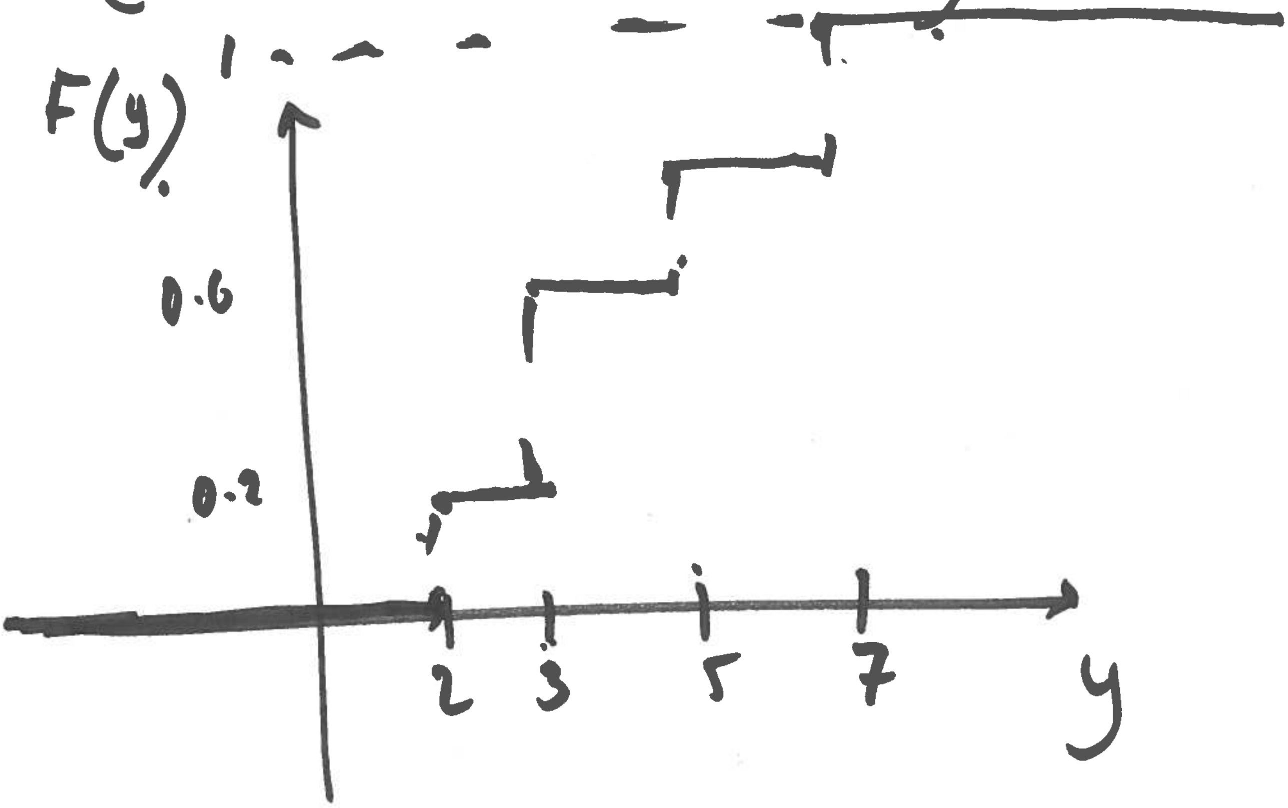
ecdf

$$F(y) = \frac{\# \text{ of obs } \leq y}{\text{Total } \# \text{ of obs.}}$$

The graph $\{y, F(y)\} \rightarrow$ E.C.D.F.

Example

$\{2, 3, 3, 5, 7\}$

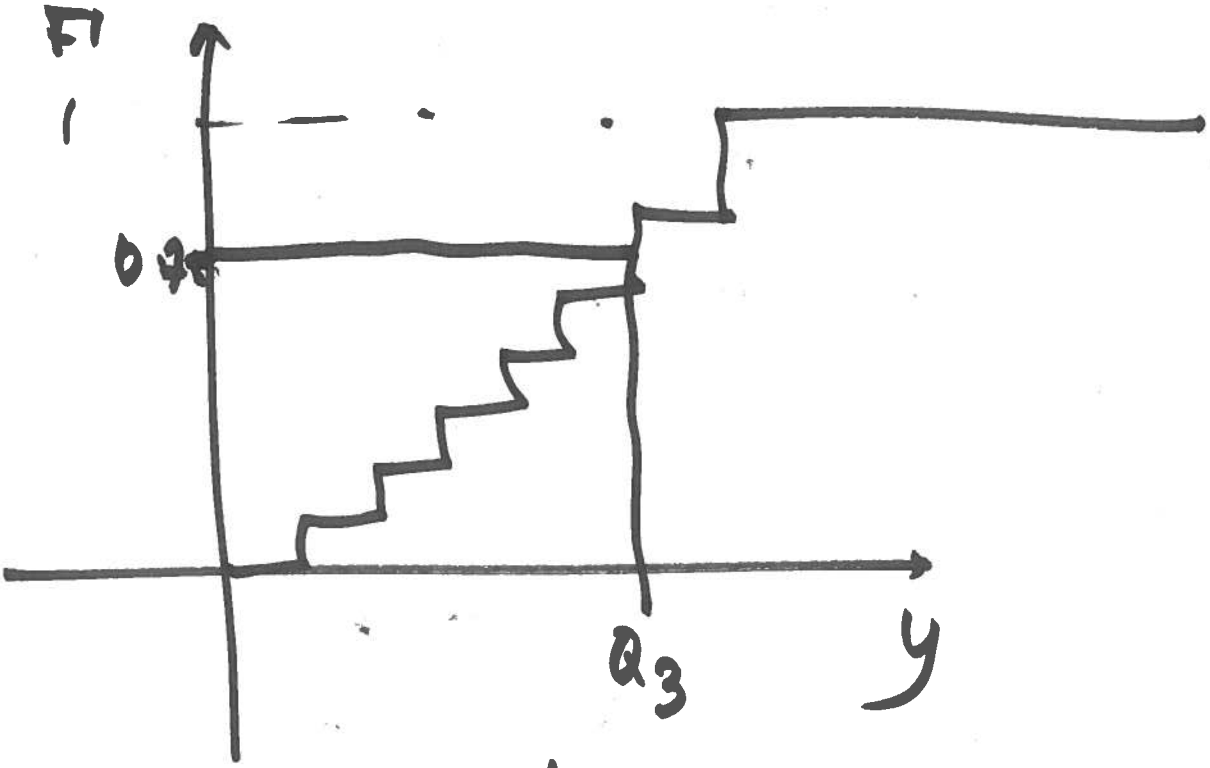


$y < 2$. $F(1) =$

$y = 2$ $F(2) = 1/5 = 0.2$

$F(2.5) = 0.2$; $F(3) =$

Empirical cdf is a step function



$0 \leq F(y) \leq 1 \quad \forall \ y$

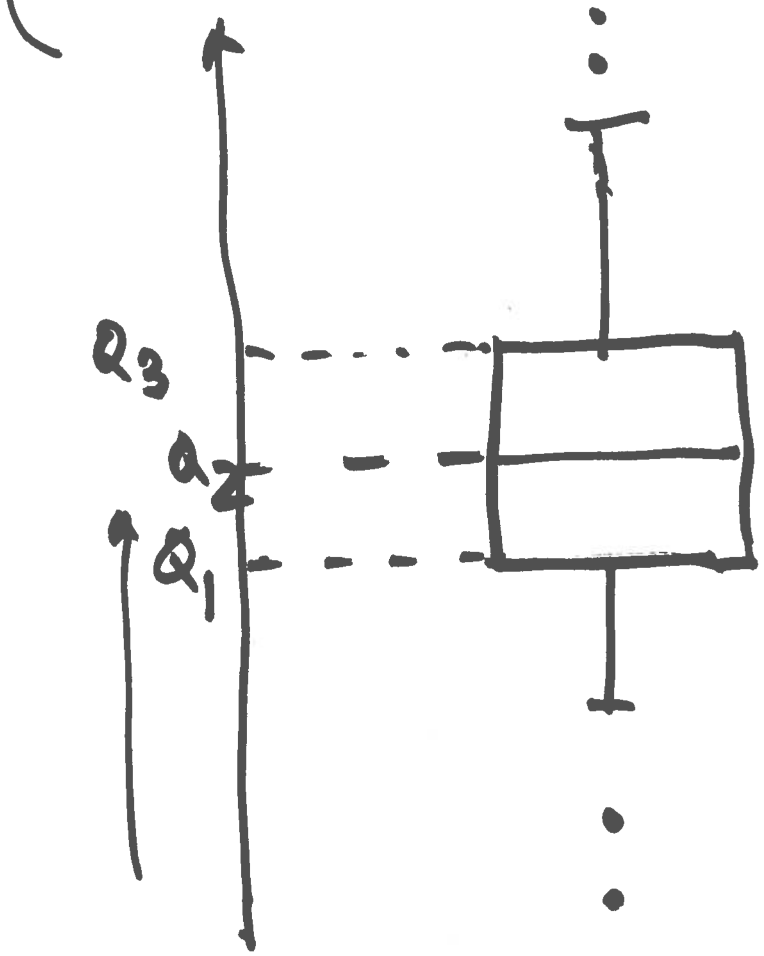
If the percentile falls on the horizont section, convention → Left most point.

The 5 # summary : $\{$Min, Q_1, Q_2, Q_3, Max$\}$

Mode = Biggest jump

Box- Plot

(Box and Whiskers plot)



Notes

(i) The width of the rectangle = immaterial

(ii) Lower end of the box = Q_1
Upper end is = Q_3, Median = Q_2
(is marked.

The whisker part:

Upper whisker stops at the maximum value of your data set $\leq Q_3 + 1.5\,IQR$

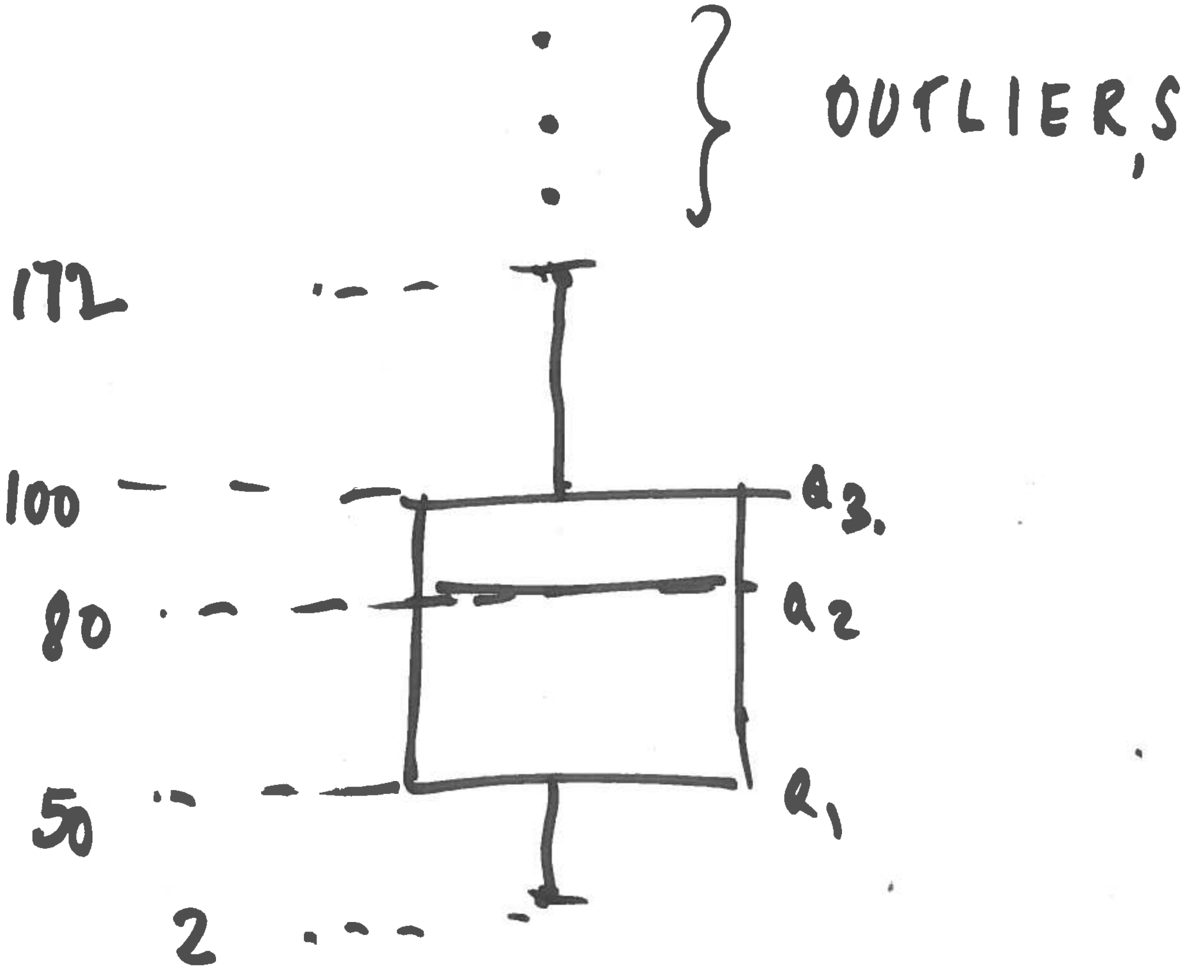
Lower whisker stops at the minimum value of the data set $\geq Q_1 - 1.5\,IQR$

Example

$\{2, 3, \ldots\ldots 172, 185, 192, 213\}$

$Q_1 = 50$ $\quad Q_3 = 100$

$Q_2 = 80$



$IQR = Q_3 - Q_1 = 100 - 50 = 50$

$1.5 \times IQR = 75$

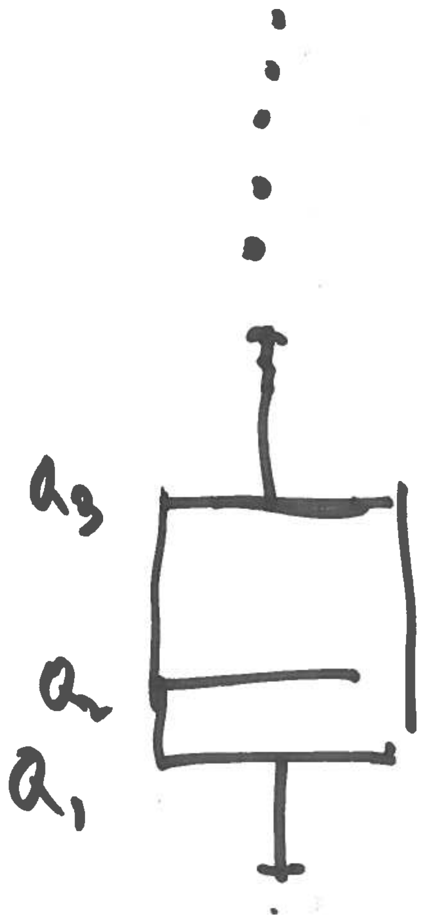
$Q_3 + 1.5\,IQR$

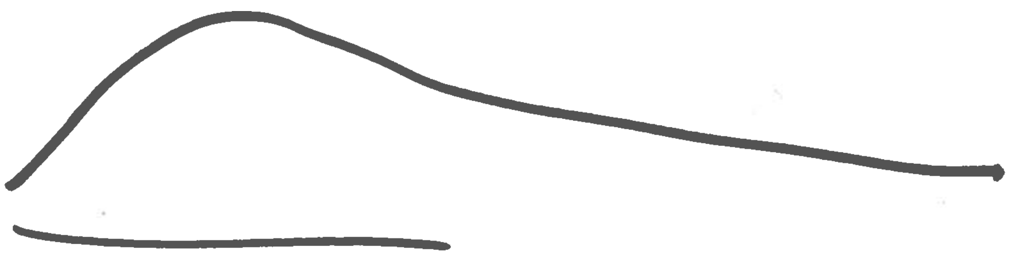
$= 175$

$Q_1 - 1.5\,IQR = 50 - 75$

$= -25$

Notes





We compare more than one data set → vertically represented

The Box-plot gives us the five # summary

It identifies each extreme observation and looks at each individually

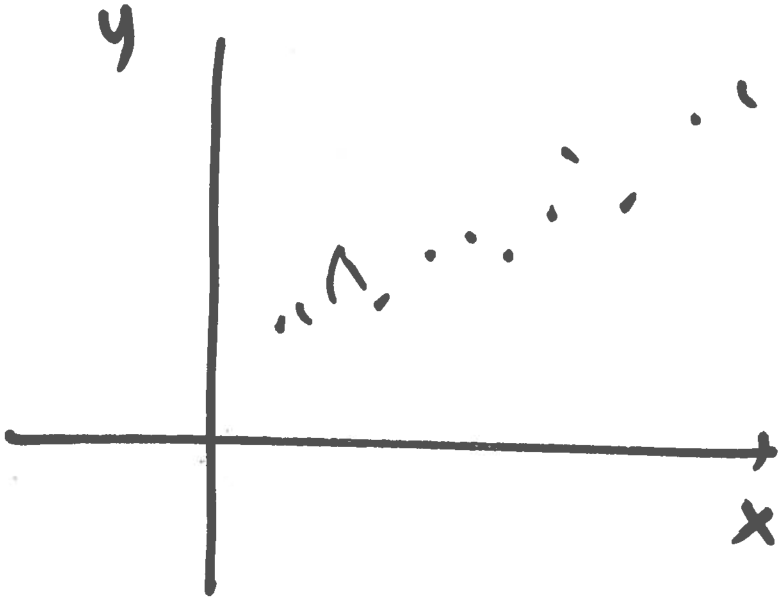
---

SCATTER PLOT : Tries to find the association between two variables x and y.

INDEPENDENT VARIABLE → Explanatory Variable

DEPENDENT → RESPONSE variate

A scatter plot: is a plot of (x, y)



Trend of some sort $\Rightarrow$ evidence of association between x and y

No obvious trend $\Rightarrow$ evidence of no association