Descriptive statistics for continuous features

Fundamentals of Machine Learning for Predictive Data Analytics,

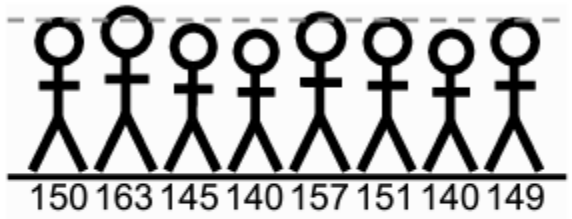
Appendix A

Central tendency

Central tendency refers to the value that is typical of the sample

Arithmetic mean (or **sample mean**, or **mean**)

$$\overline{a} = \frac{1}{n}\sum_{i=1}^{n} a_i$$



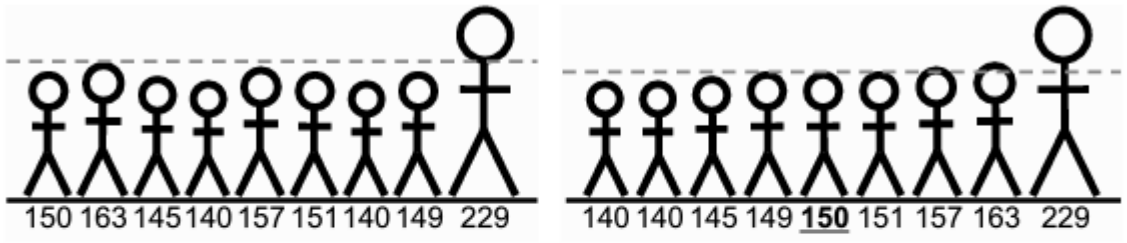
$$\overline{\text{HEIGHT}} = \frac{1}{8} \times (150 + 163 + 145 + 140 + 157 + 151 + 140 + 149)$$
$$= 149.375$$

Figure A.1

The members of a school basketball team. The height of each player is listed below the player. The dashed gray line shows the arithmetic mean of the players' heights.

Median and mode

Median: the middle value when you order the values from lowest to highest



(a) Mean (b) Median

Figure A.2

The members of the school basketball team from Figure A.1[526] with one very tall *ringer* added: (a) the dashed gray line shows the mean of the players' heights; (b) the dashed gray line shows the median of the players' heights, with the players ordered by height.

Mode: most commonly occurring value

Variation

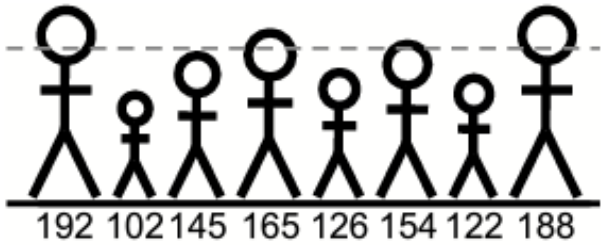


Figure A.3

The members of a rival school basketball team. Player heights are listed below each player. The dashed gray line shows the arithmetic mean of the players' heights.

Range = max – min

- Very sensitive to outliers

Range(sample_Fig1) = 163-140 = 23

Range(sample_Fig3) = 192-102=90

Variance

Variance

$$var(a) = \frac{\sum_{i=1}^{n} (a_i - \bar{a})^2}{n - 1}$$

$$var(\text{HEIGHT}) = \frac{(150 - 149.39)^2 + (163 - 149.39)^2 + \ldots + (149 - 149.39)^2}{8 - 1}$$

$$= 65.282$$

(for the sample in Fig 1)

$$var(\text{HEIGHT}) = \frac{(192 - 149.39)^2 + (102 - 149.39)^2 + \ldots + (188 - 149.39)^2}{8 - 1}$$

$$= 1{,}020.348$$

(for the sample in Fig 3)

Standard deviation

Standard deviation

$$sd(a) = \sqrt{var(a)}$$

$$= \sqrt{\frac{\sum_{i=1}^{n} (a_i - \overline{a})^2}{n-1}}$$

sd(sample_Fig1) = 8.08

sd(sample_Fig3) = 31.94

Percentiles

i^{th} percentile:

proportion of $\frac{i}{100}$ of the values in a sample are equal or lower than the i^{th} percentile

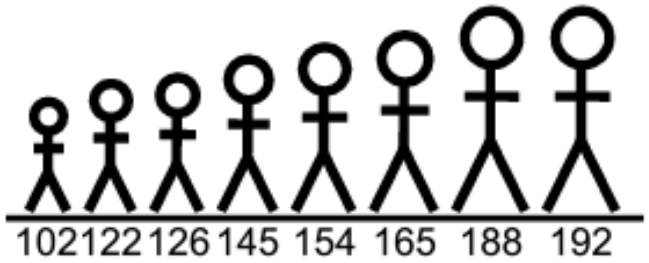


Figure A.4

The members of the rival school basketball team from Figure A.3[527] ordered by height.

1st and 3rd quartile

- Lower quartile (or 1st quartile)
  the median of the lower half of the data
- Upper quartile (3rd quartile)
  the median of the upper half of the data

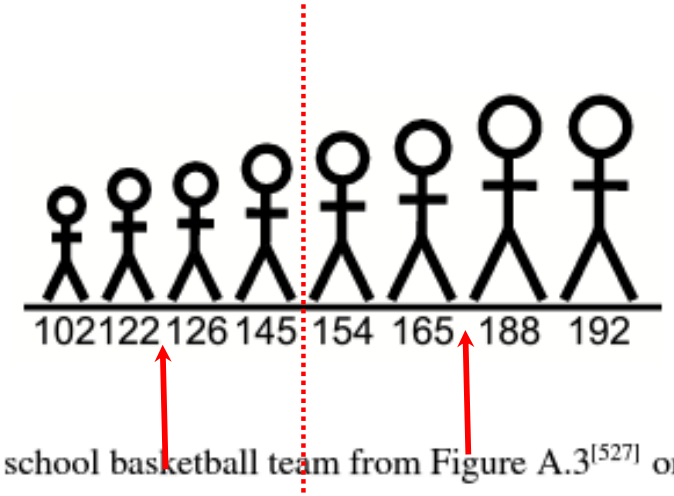


Figure A.4

The members of the rival school basketball team from Figure A.3[527] ordered by height.

Descriptive statistics for categorical features

Frequency count and proportion

Table A.1

A dataset showing the positions and monthly training expenses of a school basketball team.

ID	POSITION	TRAINING EXPENSES	ID	POSITION	TRAINING EXPENSES
1	center	56.75	11	center	550.00
2	guard	1,800.11	12	center	223.89
3	guard	1,341.03	13	center	103.23
4	forward	749.50	14	forward	758.22
5	guard	1,150.00	15	forward	430.79
6	forward	928.30	16	forward	675.11
7	center	250.90	17	guard	1,657.20
8	guard	806.15	18	guard	1,405.18
9	guard	1,209.02	19	guard	760.51
10	forward	405.72	20	forward	985.41

Table A.2

A frequency table for the POSITION feature from the school basketball team dataset in Table A.1[531].

Level	Count	Proportion
guard	8	40%
forward	7	35%
center	5	25%

Mode

Mode – most frequent level

Second mode – 2nd most frequent level

- Example

  Mode → “guard”

  2nd mode → “forward”