Data Cleaning

Problem Solving

Data

Age	%Fat
23	9.5
23	26.5
27	7.8
27	17.8
39	31.4
41	25.9
47	27.4
49	27.2
50	31.2
52	34.6
54	42.5
54	28.8
56	33.4
57	30.2
58	34.1
58	32.9
60	41.2
61	35.7

Measures of Centrality

- Mean
- Median
- Mode
- Range
- Standard Deviation
- Variance

Find the above values for *age* and *%fat*

Mean

Add all the numbers then divide by the amount of numbers

9, 3, 1, 8, 3, 6

9 + 3 + 1 + 8 + 3 + 6 = 30

30 ÷ 6 = 5

The mean is 5

Median

Order the set of numbers, the median is the middle number

9, 3, 1, 8, 3, 6

1, 3, 3, 6, 8, 9

The median is 4.5

Mode

The most common number

9, 3, 1, 8, 3, 6

The mode is 3

Range

The difference between the highest number and lowest number

9, 3, 1, 8, 3, 6

9 - 1 = 8

The range is 8

mode

The value that occurs most often in a data set.

How to determine the mode in a set of scores.

Order the scores from least to greatest.
Locate the score that occurs the most.

3, 4, 5, 5, 5, 6, 6, 7, 8, 8, 9

mode = 5

3, 4, 5, 5, 5, 6, 6, 6, 8, 8, 9

modes = 5 and 6

two modes are called bimodal
more than two modes are called multimodal

1, 2, 3, 4, 5, 6, 7, 8, 9, 10

there is no mode in this set of scores

one mode ... unimodal
two modes ... bimodal
three modes ... trimodal
more than one mode ... multimodal

Sample Variance

$$s^2 = \frac{\sum(x - \bar{x})^2}{n - 1}$$

Sample Standard Deviation

$$s = \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$$

$$Standard\ Deviation = \sqrt{Variance}$$

Answers

- For attribute ***age***

mean	46.44
median	51
modes	23,27,54,58
sd	13.21
variance	174.73
Range	38

- For attribute ***%fat***

mean	28.78
median	30.7
mode	No mode
sd	9.25
variance	85.64
Range	34.7

North Central Cancer Treatment Group (NCCTG) provides the following dataset for lung cancer prediction.

Id no	inst	time	status	age	sex	ph. ecog	ph. karno	pat. karno	meal. cal	wt.loss
1	3	306	2	74	1	1	90	100	1175	
2	3	455	2	68	1	0	90	90	1225	15
3	3	1010	1	56	1	0	90	90		15
4	5	210		57	1	1	90	60	1150	11
5	1	883	2	60	1	0	100	90		0
6	12	1022	1		1	1	50	80	513	0
7	7	310	2	68	2	2	70	60	384	10
8	11		2	71	2	2	60	80	538	1
9	1	218	2	53	1	1	70	80	825	16
10	7	166	2	61		2	70	70	271	34
11	6	170	2	57	1	1	80	80	1025	27
12	16	654	2	68	2	2	70	70		23
13	11	728	2	68	2	1	90	90		5
14	21	71	2	60	1		60	70	1225	32
15	12	567	2	57	1	1	80	70	2600	60

Exercises

North Central Cancer Treatment Group (NCCTG) provides the following dataset for lung cancer prediction.

a) Fill the missing values present in this dataset. **(Filling missing values)**

b) Create a sampled dataset of size 5 using Random sampling with and without replacements. **(Data Sampling)**

Answers Filling Missing Values

Missing values are replaced by appropriate central tendency measure

(Mean, median or mode)

Answers are marked RED

Id no	inst	time	status	age	sex	ph. ecog	ph. karno	pat. karno	meal. cal	wt.loss
1	3	306	2	74	1	1	90	100	1175	15
2	3	455	2	68	1	0	90	90	1225	15
3	3	1010	1	56	1	0	90	90	994	15
4	5	210	2	57	1	1	90	60	1150	11
5	1	883	2	60	1	0	100	90	994	0
6	12	1022	1	63	1	1	50	80	513	0
7	7	310	2	68	2	2	70	60	384	10
8	11	484	2	71	2	2	60	80	538	1
9	1	218	2	53	1	1	70	80	825	16
10	7	166	2	61	1	2	70	70	271	34
11	6	170	2	57	1	1	80	80	1025	27
12	16	654	2	68	2	2	70	70	994	23
13	11	728	2	68	2	1	90	90	994	5
14	21	71	2	60	1	1	60	70	1225	32
15	12	567	2	57	1	1	80	70	2600	60