Descriptive Statistics

19 December 2022 08:06

Statistical methods

Descriptive — Inferential statistic

DA: ① read ② Analysis (80%)

ML: ① read data ② clean data ③ preprocess data ④ model building

Statistical methods

80% → DA analysis; 20% → ML getting result

Data past → Analysis; Data present → ML → future pred

Stats :- Representation of your data in terms of graph or Tables and applying some statistical methods to get interpretations / results about your data

Descriptive Stats: Ds Analysation of past data and get the pattern for past data which are useful for present / future data.

Descriptive → Analysation, patterns

future → prediction → ML / Hypothesis → Inferential

Descriptive :-

- Measures of Central Tendency
- Measures of Dispersion / Deviation
- Measures of shape
- measures of relation

x: 10, 11, 12, 13

$\frac{10+11+12+13}{4} = \square$

Central Tendency:

① mean :- Sum of all obsv / no of the obsv ($\bar{x}$) / (μ) $\bar{x} = \frac{\sum x}{n}$ (summation)

② median :- middle value / which is divided into two equal parts of data

③ mode

median: 50% | 50%

① 7, 15, 13, 6, 4, 2, 10,

mean = $\frac{7+15+13+6+4+2+10}{7}$ = 57/7 = 8.14

median =

median :- 2, 4, 6, ⑦, 10, 13, 15 ⇒ m = 7

② 7, 15, 13, 6, 4, 2, 10, (100)

mean = 157/8 = 19.6

n = 8

2, 4, 6, ⑦, ⑩, 13, 15, 100

median = $\frac{\frac{n}{2}^{th} + (\frac{n}{2}+1)^{th}}{2}$

$= \frac{(\frac{8}{2})^{th} + (\frac{8}{2}+1)^{th}}{2}$

$= \frac{4^{th} + 5^{th}}{2}$

median = 8.5

median = 7

mean = 19.6

mean = 8.14

19.6 > 8.5

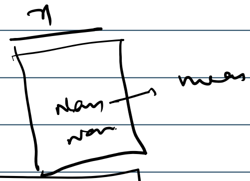
mean = 8.5 (boxed)

19.6 > 8.5 = (7 + 8)/2

= (7 + 10)/2 = 17/2 = 8.5

Note:- mean is effected by the outliers but not median

CT:- ① Imputation of the ~~data~~ null value
② Identifying the distn

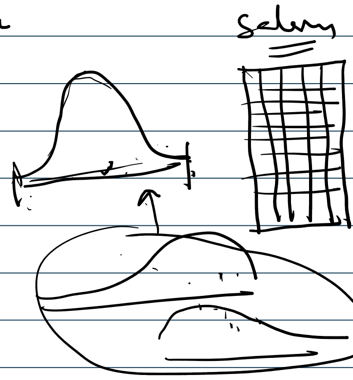


mean > median

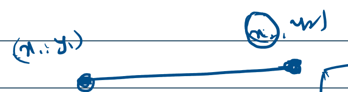
mode:- High concentration / Highly repeated value

2, 4, 5, 1, 2, 4, 1, 2, 2 → mode: 2 – 4 times
1 – 2 times

Numerical	Categorical
① Imputation of null value we use mean / median	① Imputation of null value we use mode
②	



Job role
value_count
Data – 1000
DA – 1000
DS – 890
DE –
MLE –



$\sqrt{(x_2-x_1)^2+(y_2-y_1)^2}$

Measures of Dispersion,

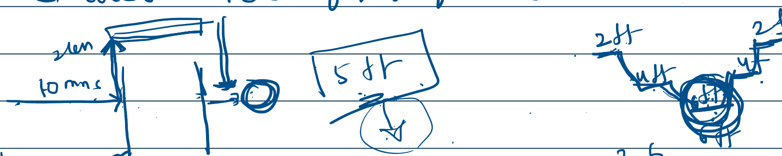
① Range :- Max − min (H − L)
② SD – Standard deviation : Distance b/w sample data
③ variance

SD:- Distance b/w original point / observation to mean

$$SD = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n}} = \sigma \text{ (sigma)}$$ (Root mean squared Error) RMSE

var:- $\sigma^2 = \frac{\Sigma(x-\bar{x})^2}{n}$ (mean squared error)

*** SD is also identifying of the distn



$\frac{2+4+6+4+2}{5} = 18/5 = 3.6$ → mean

x	$(x-\bar{x})$	$(x-\bar{x})^2$
2	2-3.6 = -1.6	2.56
4	4-3.6 = 0.4	0.16
6	6-3.6 = 2.4	5.76

$\sqrt{\frac{8.1}{5}}$ =

2ft, 4ft, 6ft, 4ft

SD = 1.4

0.1
0.2

4	$4-3.6 = 0.4$	0.16
6	$6-3.6 = 2.4$	4.86
4	$4-3.6 = 0.4$	0.16
2	$2-3.6 = -1.6$	2.56
		10.30

$\sigma = \sqrt{\frac{(x-\bar{x})^2}{n}}$

$\sigma = \sqrt{\frac{10.3}{5}} = \sqrt{2.06}$

$SD = 1.4$ ← mean = 3.6

$SD = 0.4$

$\frac{16}{5} = 3.2$

x	$(x-\bar{x})$	$(x-\bar{x})^2$	$(x-\bar{x})^2$
2	$2-3.2 = -1.2$		1.44
4	$4-3.2 = 0.8$		0.64
4	$4-3.2 = 0.8$		0.64
4	$4-3.2 = 0.8$		0.64
2	$2-3.2 = -1.2$		1.44
			4.80

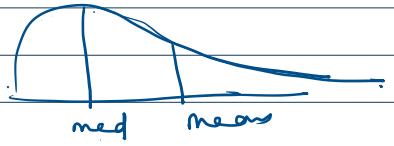
$\sigma = \sqrt{\frac{4.80}{5}} = \sqrt{0.96}$

$SD = 0.98$

```
diamond_df['price'].describe()
```

```
count    53940.000000
mean      3932.799722
std       3989.439738
min        326.000000
25%        950.000000
50%       2401.000000
75%       5324.250000
max      18823.000000
Name: price, dtype: float64
```

mean > med positive skewness



SD > mean