

Exploring Data through Summary Statistics

After this video you will be able to..

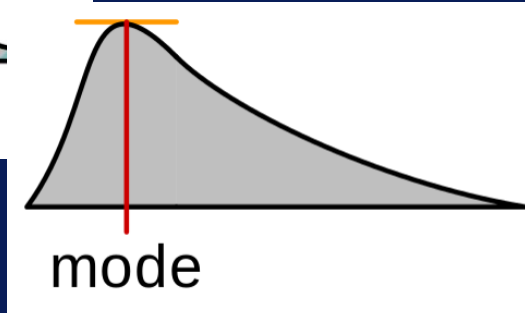
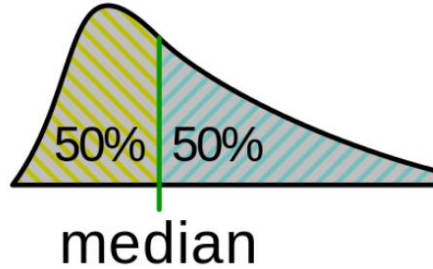
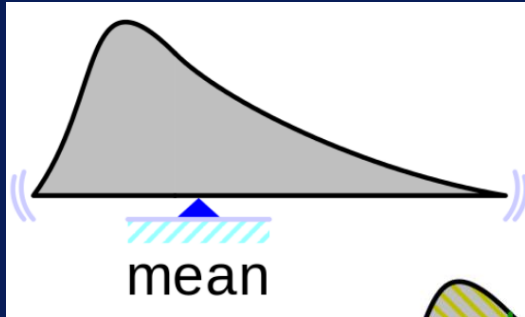
- Define what a summary statistic is
- List three common summary statistics
- Explain how summary statistics are useful in exploring data

What are summary statistics?

- Quantities that summarize and describe a set of data values
- Measures of
 - Location: mean, median
 - Spread: standard deviation
 - Shape: skewness

Measures of Location

Describe central or typical value of dataset



Measures of Location - Example

Age	Age (sorted)
35	21
42	22
78	35
22	42
56	42
50	50
42	56
78	78
21	78
87	87

Mean = 51.1

Median = $(42+50)/2 = 46$

Mode = 42 & 78

Measures of Spread

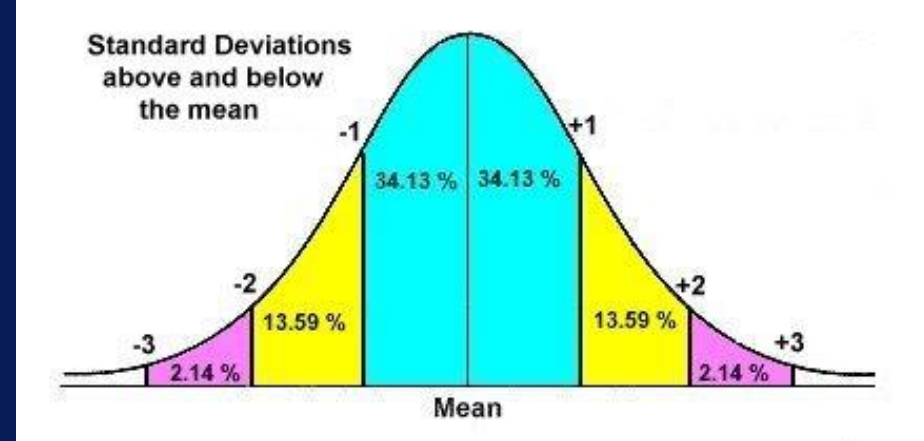
Describe how dispersed or varied data is

minimum maximum

standard deviation

variation

range



Measures of Spread – Example

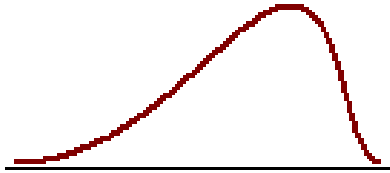
Age	Age (sorted)
35	21
42	22
78	35
22	42
56	42
50	50
42	56
78	78
21	78
87	87

$$\text{Range} = 87 - 21 = 66$$

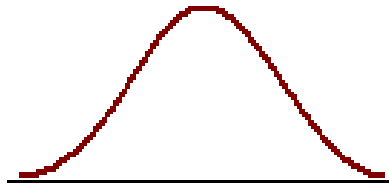
$$\text{Variance} = 548.767$$

$$\text{Standard deviation} = 23.426$$

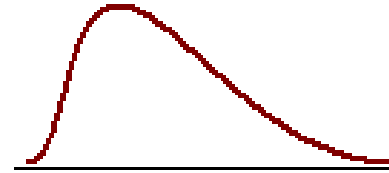
Measures of Shape



Negatively skewed distribution
or Skewed to the left
Skewness < 0



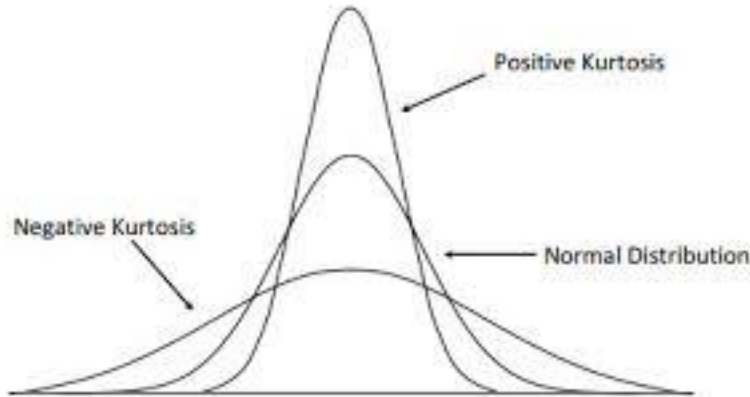
Normal distribution
Symmetrical
Skewness $= 0$



Positively skewed distribution
or Skewed to the right
Skewness > 0

skewness

kurtosis



Measures of Shape – Example

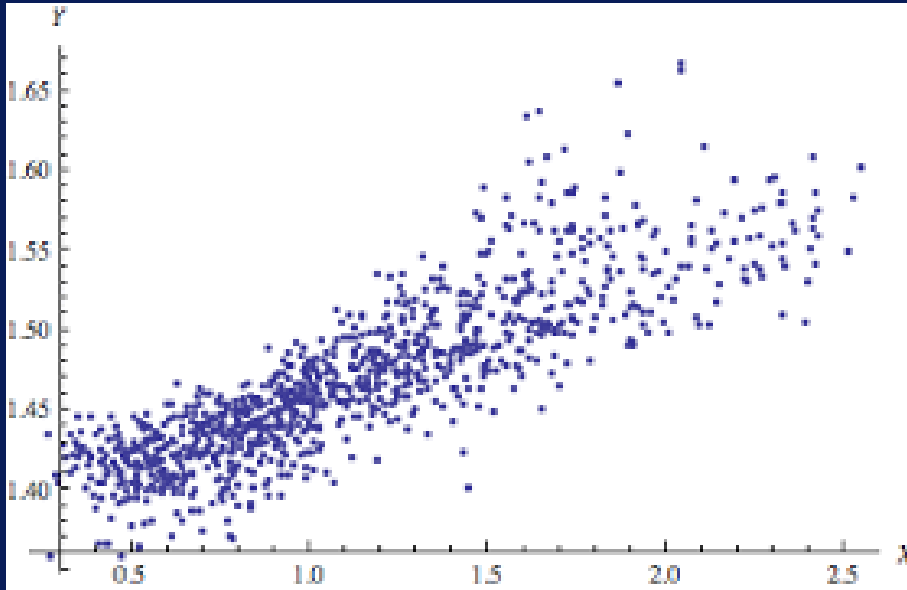
Age
35
42
78
22
56
50
42
78
21
87

Skewness = 0.2995

Kurtosis = -1.2028

Measures of Dependence

Describe relationship between variables



correlation

Measures of Dependence – Example

Height	Weight
180	68
153	70
204	84
133	44
208	81
142	53
122	40
168	50
175	64
200	72

Correlation = 0.8906

Statistics on Categorical Variables

Describe number of categories and
frequency of each category

Color/Pet	White	Brown	Black	Orange	Total
Dog	34	44	32	0	110
Cat	25	2	43	0	70
Fish	1	0	5	33	39
Total	60	46	80	33	219

contingency table

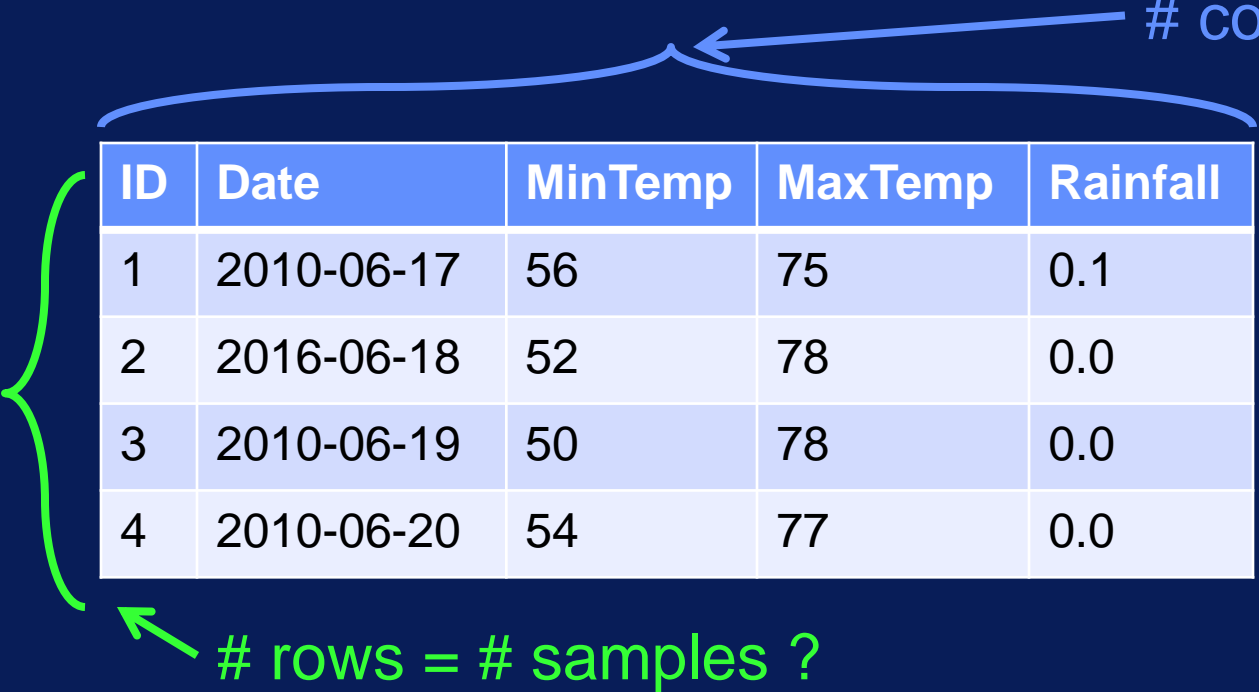
Contingency Table - Example

Color/Pet	White	Brown	Black	Orange	Total
Dog	34	44	32	0	110
Cat	25	2	43	0	70
Fish	1	0	5	33	39
Total	60	46	80	33	219

Check Dimensions

- Check number of rows and columns

columns = # variables ?



ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	56	75	0.1
2	2016-06-18	52	78	0.0
3	2010-06-19	50	78	0.0
4	2010-06-20	54	77	0.0

rows = # samples ?

Check Values

- Check values in some samples

Should temperature values in F or C?

Is this correct?

ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	56	24	0.1
2	2016-06-18	52	26	3,678.9
3	2010-06-19	50	26	0.0
4	2010-06-20	54	25	0.0

Is this date or timestamp?

Check Missing Values

ID	Date	MinTemp	MaxTemp	Rainfall
1	2010-06-17	56	75	--
2	2016-06-18	52	78	--
3	2010-06-19	--	78	0.1
4	2010-06-20	54	77	--

Does feature
have mostly
missing values?



How many samples have
missing values?

Summary Statistics

- **Measures of**
 - Location, spread, shape, dependence
- **Contingency table**
 - For categorical variables
- **Data validation**
 - Dimensions, missing values