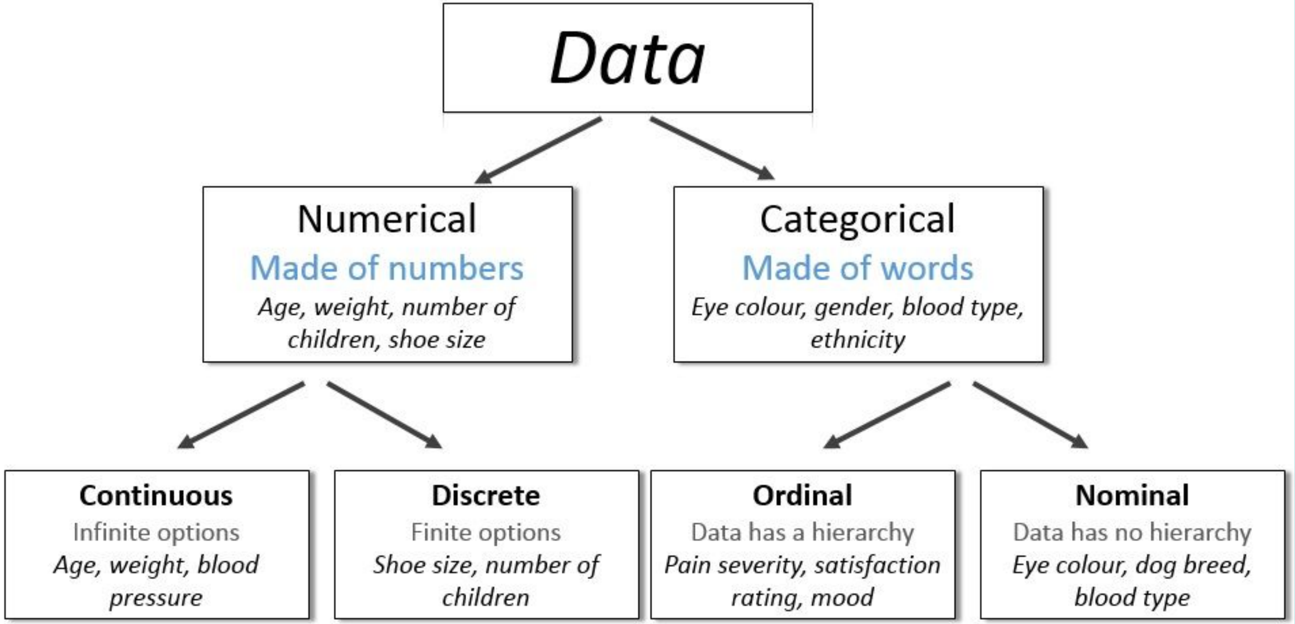
WELCOME TO

TECH I.S.

Descriptive Statistics

Types of Data



Data

Numerical

Made of numbers

Age, weight, number of children, shoe size

Categorical

Made of words

Eye colour, gender, blood type, ethnicity

Continuous

Infinite options

Age, weight, blood pressure

Discrete

Finite options

Shoe size, number of children

Ordinal

Data has a hierarchy

Pain severity, satisfaction rating, mood

Nominal

Data has no hierarchy

Eye colour, dog breed, blood type

Measure of Center (Central Tendency)

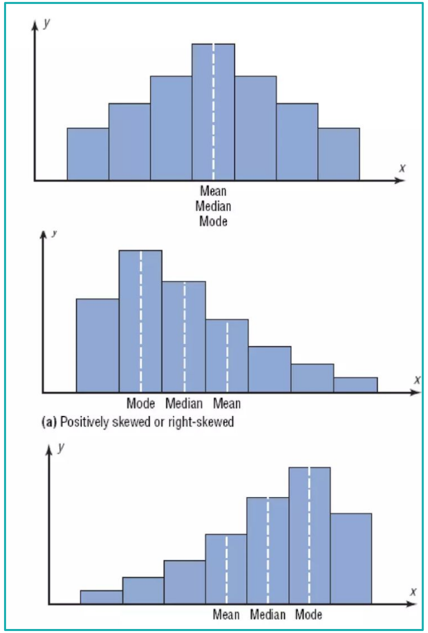
A measure of center is a value at the center or middle of a data set.

Mean: $\bar{x} = \frac{\sum x}{n}$,

Median: The **middle value** of ranked data

Mode: The value(s) that occur(s) with the greatest frequency.

Midrange: $Mr = \frac{Min+Max}{2}$



(a) Positively skewed or right-skewed

Measures of Variability

Variability refers to the spread of the values within a distribution.

- Range:
  - Difference Between the highest and lowest scores.
- Variance:
  - The degree of spread within the distribution (the larger the spread, the larger the variance).
- Standard Deviation:
  - A measure of how the average score deviates or spreads away from the mean (defined as the square root of the variance).

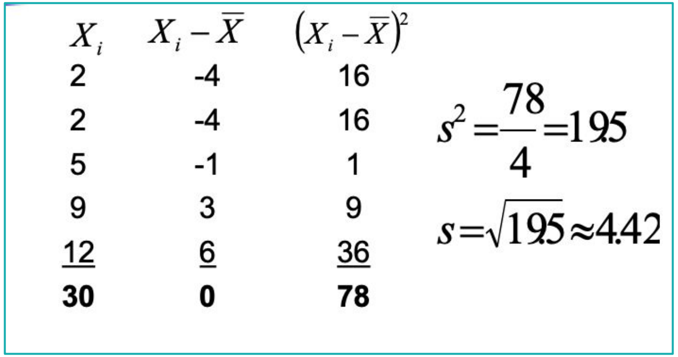
Sample Variance and Standard Deviation

Sample variance (s^2) is a measure of the degree to which the numbers are spread out.

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

Sample Standard deviation measures the spread of a data in terms of distance between each data point and mean.

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



X_i	$X_i - \bar{X}$	$(X_i - \bar{X})^2$
2	-4	16
2	-4	16
5	-1	1
9	3	9
12	6	36
30	**0**	**78**

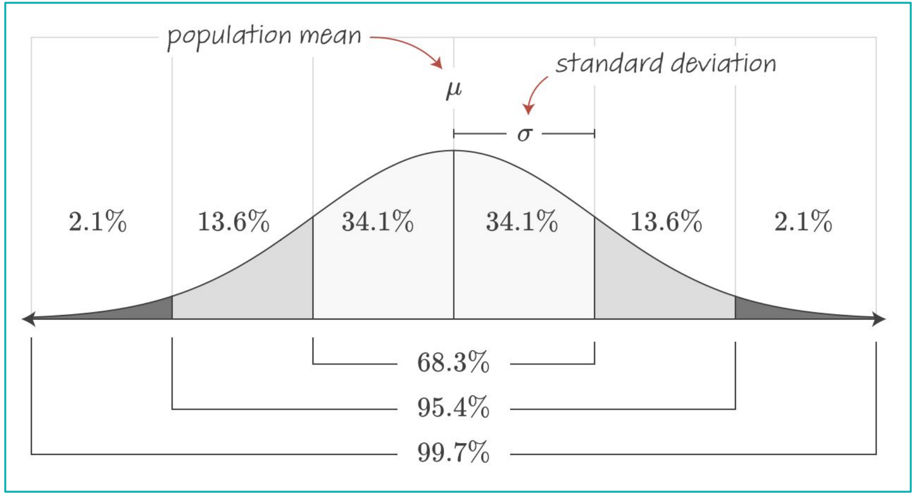
$$s^2 = \frac{78}{4} = 19.5$$

$$s = \sqrt{19.5} \approx 4.42$$

Interpreting the Standard Deviation

Empirical Rule (68-95-99 rule):

- 68% data lies within 1 standard deviation of mean.
- 95% data lies within 2 standard deviations of mean.
- 99.7% data lies within 3 standard deviations of mean.



Measures of Relative Standing

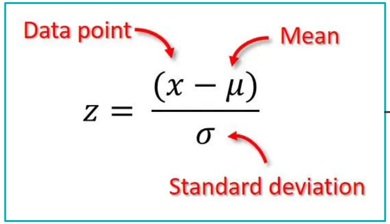
Measures of relative standing are numbers showing the location of data values relative to the other values within the same data set.

1. Z-score
2. Percentile
3. Quartile
4. Boxplot

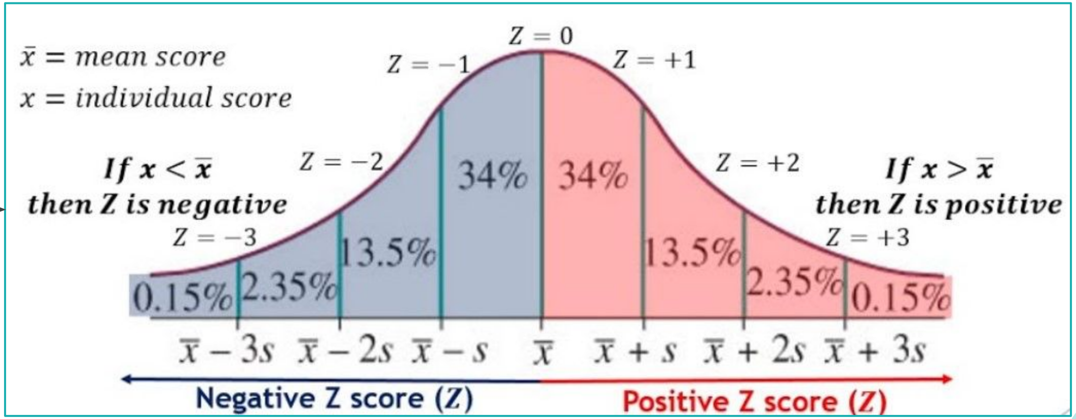
Z-Score

Z-Score: The number of Standard Deviations from the Mean.

- If Z > 0 then X_i is greater than mean
- If Z < 0 then X_i is less than mean



$$z = \frac{(x - \mu)}{\sigma}$$

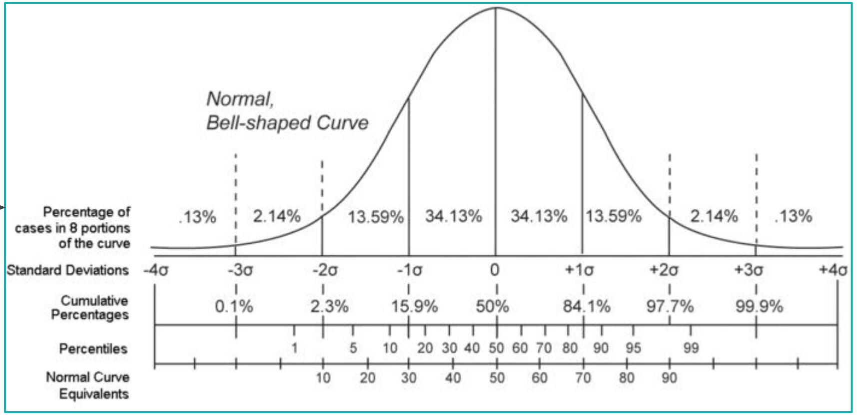


Percentile Rank

<u>Percentile rank (PR) refers to the position within a group that a person with a particular score is at.</u>

- A person with percentile rank of means that he /she scored better than 70 percent of the group.
- PR 50 means person scored same or better than 50 percent of group.

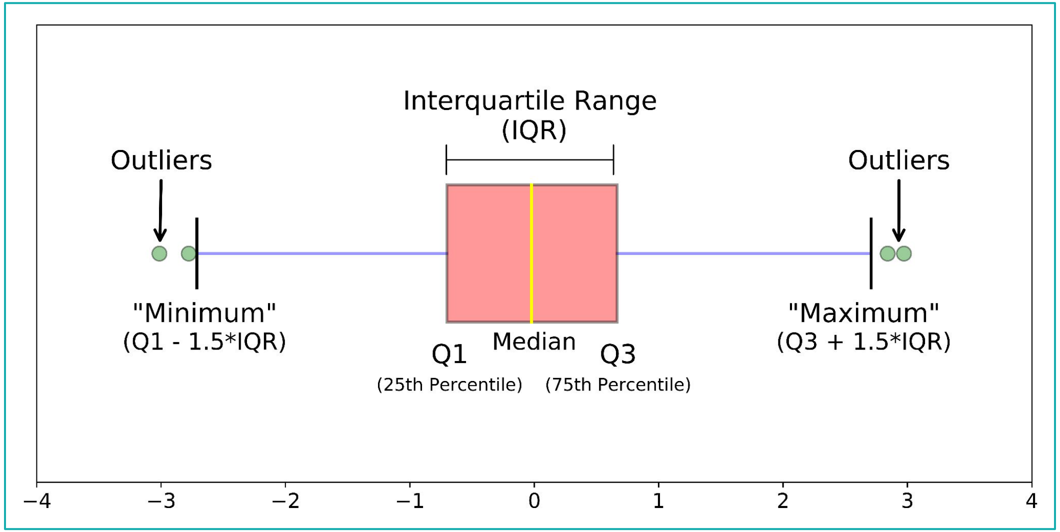
$$PR = \frac{CF - (0.5 \times F)}{N} \times 100,$$



Box Plots

A box plot is a graphical display, based on quartiles, that helps to picture a set of data.

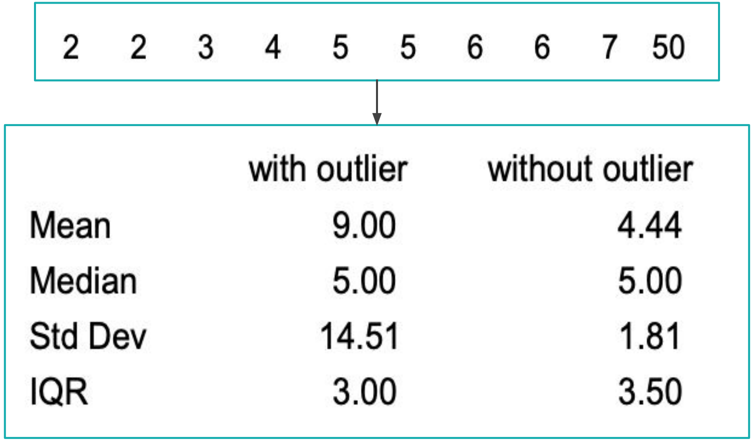
Five pieces of data are needed to construct a box:



Outliers

An outlier is data point that is far removed from the other entries in the data set.

- Mistakes made in recording data
- Data that don’t belong in population
- True rare events



2 2 3 4 5 5 6 6 7 50

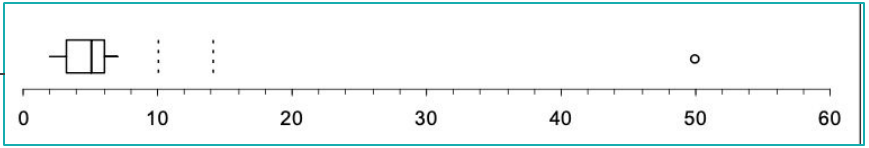
	with outlier	without outlier
Mean	9.00	4.44
Median	5.00	5.00
Std Dev	14.51	1.81
IQR	3.00	3.50

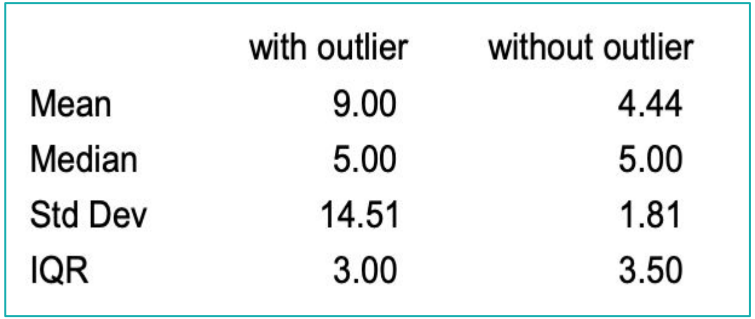
Using Box Plot to find Outliers

The “box” is the region between the 1st and 3rd quartiles.

- Possible outliers are more than 1.5 IQR from the box (inner fence)
- Probable outliers are more than 3*IQR from the box (outer fence)
- In the box plot below, the dotted lines represent the “fences” that are 1.5 and 3 IQR from the box. See how the data point 50 is well outside the outer fence and therefore an almost certain outlier.

2 2 3 4 5 5 6 6 7 50



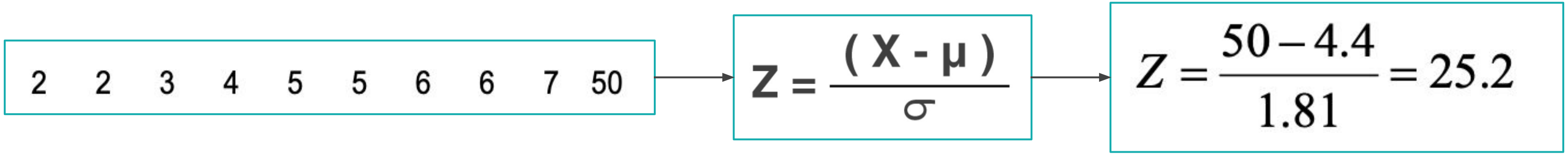


	with outlier	without outlier
Mean	9.00	4.44
Median	5.00	5.00
Std Dev	14.51	1.81
IQR	3.00	3.50

Using Z-score to detect outliers

<u>The Z-score as be used to detect outliers:</u>

- **Calculate the mean and standard deviation without the suspected outlier.**
- **Calculate the Z-score of the suspected outlier .**
- **If the Z-score is more than 3 or less than -3, that data point is a probable outlier.**



2 2 3 4 5 5 6 6 7 50

$$Z = \frac{(X - \mu)}{\sigma}$$

$$Z = \frac{50 - 4.4}{1.81} = 25.2$$

Outliers – Remove or Not?

Remove or not remove, there is no clear answer.

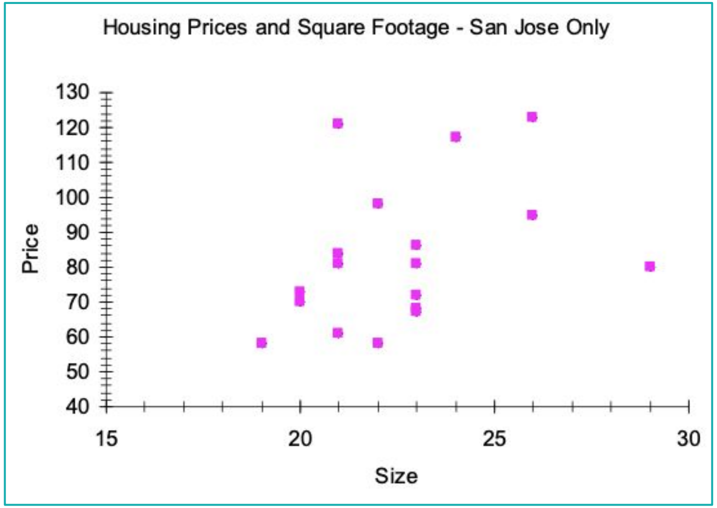
- For some populations, outliers don't dramatically change the overall statistical analysis. Example: the tallest person in the world will not dramatically change the mean height of 10000 people.

- However, for some populations, a single outlier will have a dramatic effect on statistical analysis (called "Black Swan" by Nicholas Taleb) and inferential statistics may be invalid in analyzing these populations. Example: the richest person in the world will dramatically change the mean wealth of 10000 people.

Bivariate Data

<u>**Ordered numeric pairs (X,Y) where both values are numeric**</u>

<u>Example: Housing Data</u>: Let X-axis be Square Footage and Y-axis be Price



Correlation Analysis

A group of statistical techniques used to measure the strength of the relationship (correlation) between two variables.

- **Scatter Plot: A chart that portrays the relationship between the two variables of interest.**
  - **Dependent Variable: The variable that is being predicted or estimated. “Effect”**
  - **Independent Variable: The variable that provides the basis for estimation. It is the predictor variable.**

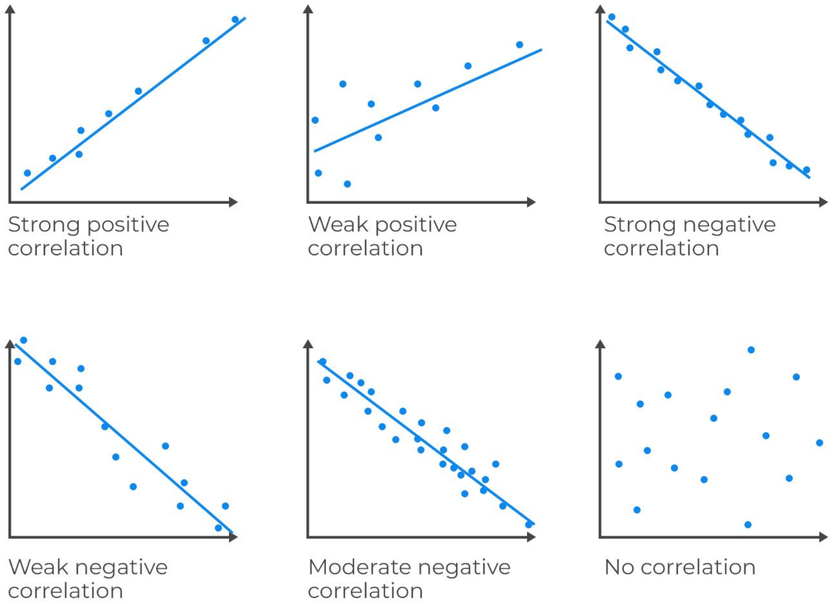
The Coefficient of Correlation

The Coefficient of Correlation (r) is a measure of the strength of the relationship between two variables.

- It requires interval or ratio-scaled data (variables).
- It can range from -1 to 1.
- Values of -1 or 1 indicate perfect and strong correlation.
- Values close to 0 indicate weak correlation.
- Negative values indicate an inverse relationship and positive values indicate a direct relationship.

Types of Correlation

(INDICATES THE RELATIONSHIP BETWEEN OF SETS OF DATA)



Strong positive correlation

Weak positive correlation

Strong negative correlation

Weak negative correlation

Moderate negative correlation

No correlation

Much obliged.

TECH I.S.