INSAID

Statistics basics for Data Science Cheat Sheet

MEASURES OF CENTRAL TENDENCY

Property	Formula	What to remember
Mean	$x = \Sigma x / n$	Sensitive to extreme values. More useful when data is symmetrical.
Median	For n = odd; Median = $(n+1)/2$ For n = even; Median = Avg. of $n/2$ and $(n+1)/2$	Not sensitive to extreme values. More useful when data is skewed.
Mode	Highest repeating value in dataset	Only measure of centre. Appropriate for categorical data.

MEASURES OF VARIATION

Property	Formula	What to remember
Sample Variance	$S^2 = \Sigma(x - \tilde{x}) / (n - 1)$	Not often used.
Sample Standard Variation	$S = \sqrt{[\Sigma(x - \tilde{x}) / (n - 1)]}$	Square root of variance. Sensitive to extreme values. Commonly used.
Interquartile Range(IQR)	IQR = 3rd quartile - 1st quartile	Less sensitive to extreme values
Range	Highest Value - Lowest Value	Not often used. Highly sensitive to unusual values. Easy to compute.

MEASURES OF RELATIVE POSITION

Property	Formula	What to remember
Percentile	Data is divided into 100 equal parts by increasing order.	For applying normal distributions
Quartile	Data is divided into 4 equal parts. For ex. Q3(third is the value greater than ¾ of others.	Used to compute IQR
Z Score	$z = (x - \tilde{x}) / s$; to find value of some observation(x) when z score is known	Measures distance from mean in terms of standard deviation