Terminology & Definitions

EDA: EXPLORATORY DATA ANALYSIS

Symbols

- S^2 = Sample variance
- S = Sample standard deviation
- σ^2 = Population (true or theoretical) variance
- σ = Population standard deviation
- $\bar{X}$ = Sample mean
- μ = Population mean
- IQR = interquartile range (middle 50%)

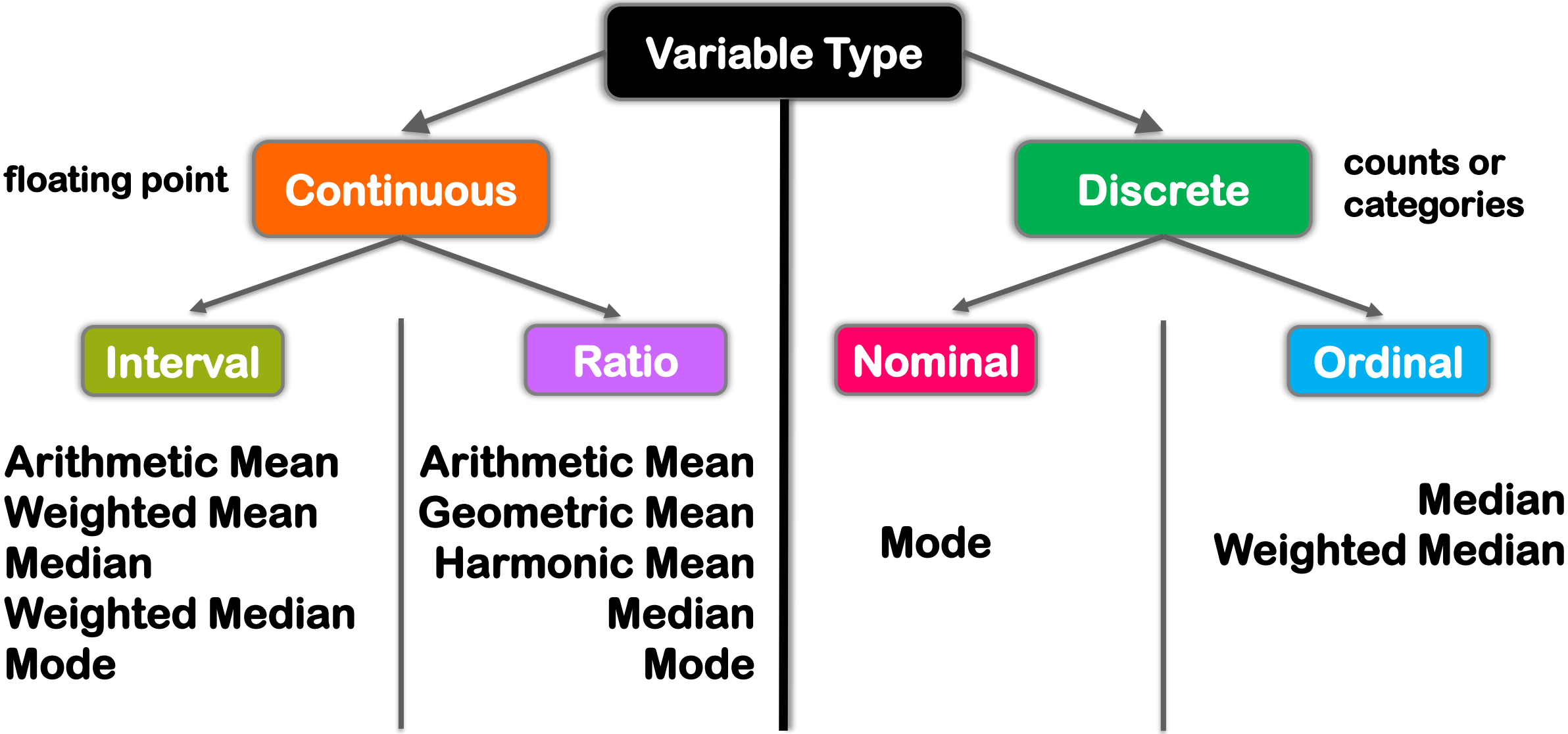
Tests: Central Tendency and Variability

Comparison	Groups	Normal or Almost Normal	Not Normal	Binomial (Proportions)	Variances
Compare data within one group to a standard or target value	1	One sample t-test	Wilcoxon Rank-Sum test	One proportion z-test (or exact Binomial test)	Chi-square for one variance
Compare data within two unpaired groups	2	Two sample t-test	Mann Whitney Wilcoxon Rank-Sum test (or U-test)	Two proportions z-test, Chi-square test of independence (or Fisher's exact test if counts in cells <5)	F-test for homogeneity of variances
Compare two paired groups	2	Paired t-test	Wilcoxon Rank-Sum test	McNemar's test	Bonett's test
Compare data among many groups	>2	One-Way ANOVA	Kruskal-Wallis test	Chi-square test of independence (or Fisher's exact test)	Levene's test or Bartlett's test for normal data

Centrality Tendency: Location Measures

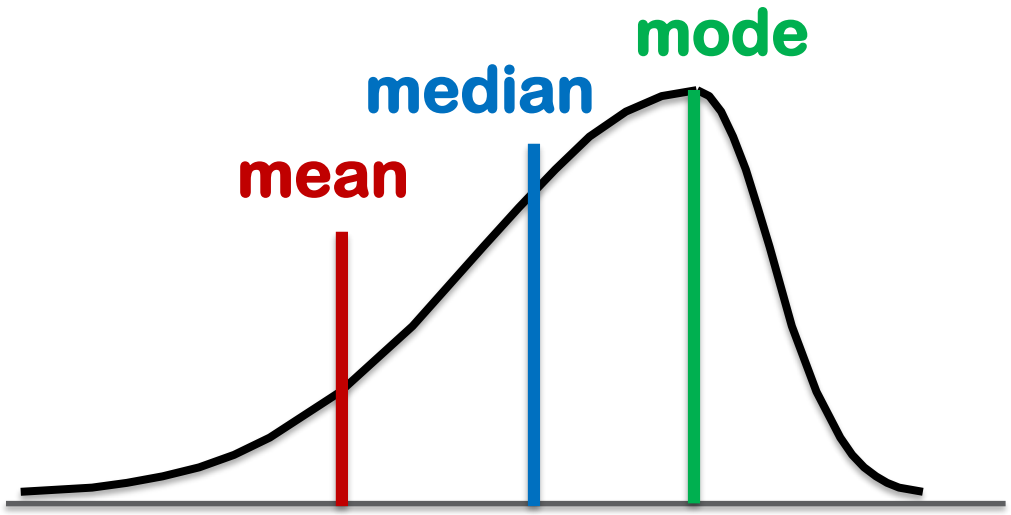
Measure	Description	Synonyms
Mean	sum of all values divided by the number of values	arithmetic mean, average
Weighted Mean	sum of all values times a weight divided by the sum of the weights	weighted average
Median	value such that one-half of the data lies above and the other-half lies below	50^{th} percentile
Weighted Median	value such that the sum of the weights is equal for the lower and upper halves of the sorted list of data values	
Trimmed Mean	average of all values after dropping a fixed number of extreme values	truncated mean
Robust	not sensitive to extreme values, or outliers	resistant
Outlier	data value that is very different from most of the data	extreme value
Mode	the most frequently observed value in the data	
Geometric Mean	characteristic of the average growth rate between positive values	
Harmonic Mean	characteristic of an average rate	

Centrality: Continuous vs. Discrete Variable



Left-Skewed Location Measures

mean is to the left of the **median**

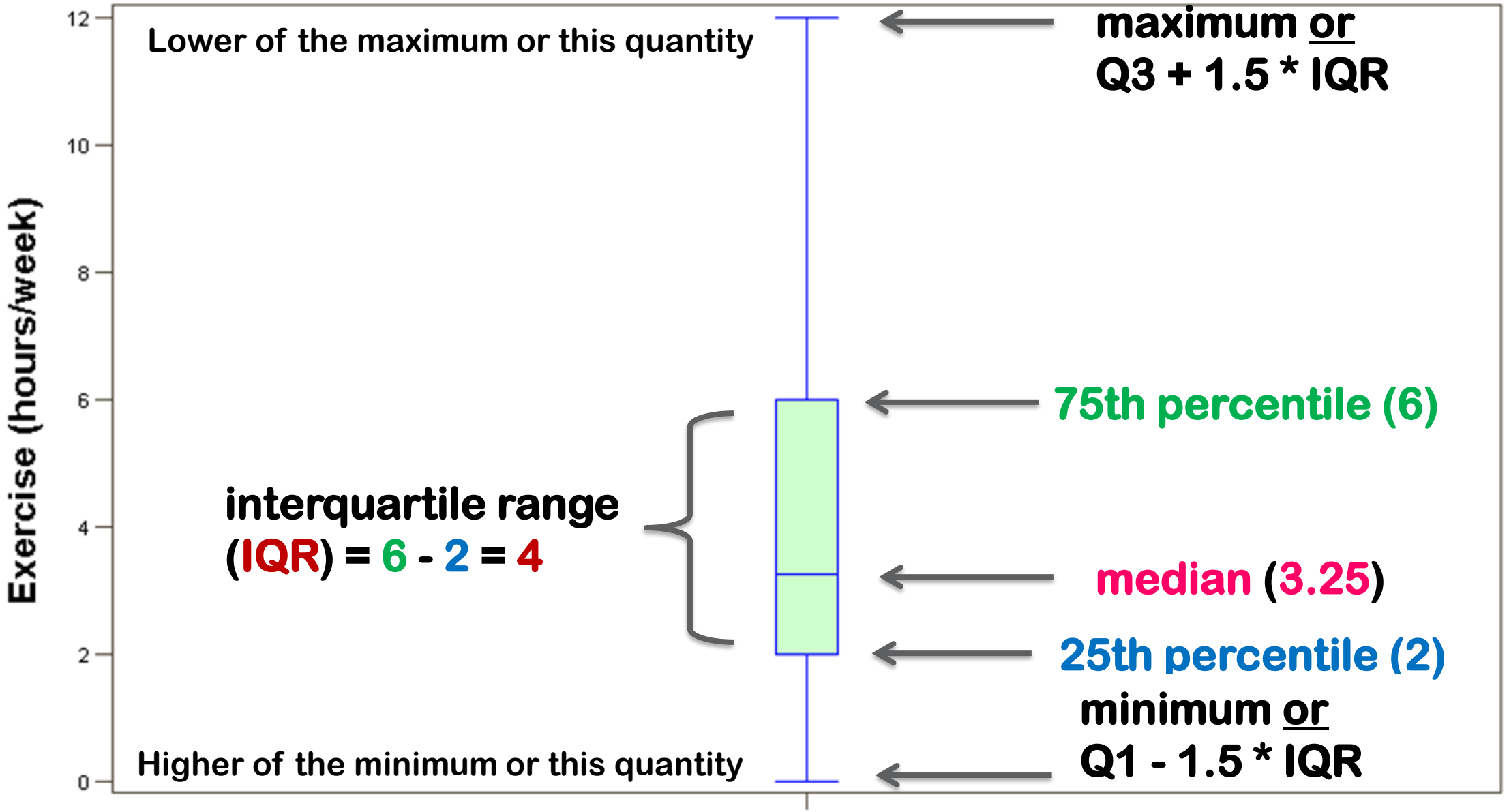


left-skewed distribution

Measures of Variability

Measure	Description	Synonyms
Range	difference between the largest and the smallest observations	
Variance	average distance from the mean	mean-squared-error
Standard deviation	average spread around the mean = square root of the variance	Euclidean norm, L2-norm
Mean/median absolute deviation	mean/median of the absolute value of the deviations from the mean	Manhattan norm, L1-norm
Percentile	value such that P percent of the values take on this value or less and (100-P) percent take more	quantile
Inter-quartile range (IQR)	difference between the 75^{th} percentile and the 25^{th} percentile of the sorted data values	IQR
Deviance	difference between the observed value and the estimated location	errors, residuals
Order statistics	metrics based on the data values sorted from smallest to biggest	ranks

IQR: Interquartile Range



Bivariate Analysis: Correlation, Covariance, Scatter

Measure	Description	Comments
Correlation coefficient	metric that measures the extent to which two numeric variables are associated with one another	ranges from -1 to +1 cor (x,y)
Correlation matrix	table where variables are shown on both rows and columns; cell values are correlations between variables	corrplot::corrplot(), corrplot::corrplot.mixed(),
Correlation test	test that measures statistical significance of the correlation coefficient	cor.test()
Scatterplot	plot in which the x-axis is the value of one variable, and the y-axis is the value of the other variable	car::scatterplotMatrix(), gpairs::gpairs()
Centering	subtracting the mean from original values	xc = x – mean(x)
Covariance	average association between two centered variables	cov (x,y)
Correlation	covariance scaled by sd(x) * sd(y); Pearson correlation	cor (x,y) = cov(x,y) / (sd(x)*sd(y))
Z-score, z	centered variable divided by its sd(x): z = xc /sd(x)	mean (z) = 0, sd(z) = 1

Bivariate Analysis: Categorical & Quantitative Vars

Measure	Description	Comments
Violin plot	similar to a boxplot but showing the density estimate	
Scatterplot matrix	plot in which the x-axis is the value of one variable, and the y-axis is the value of the other variable; but grouped-by values of the categorical variable	car::scatterplotMatrix(), Note: use pipe (|) for response variable, group-by=TRUE
Boxplot	plot to visualize distribution of data grouped by the values of the categorical variable	