Week 2. Measuring Center and Spread of Data

Categorial Variables: Compute **frequencies** of each data value (# occurrences of value).

Mode = "most frequent value"

Proportion = # occurrences of value / # data points

Numerical Variables: Two possibilities for "*center*" and "*spread*" (one from **ordering** and one from **arithmetic**)!

Center

Median = "middle value"
value of average datapoint

Order data and choose middle value

→ Not changed by outliers ("robust")
→ Difficult to compute and analyze

An **outlier** is a datapoint "far" from *all* others

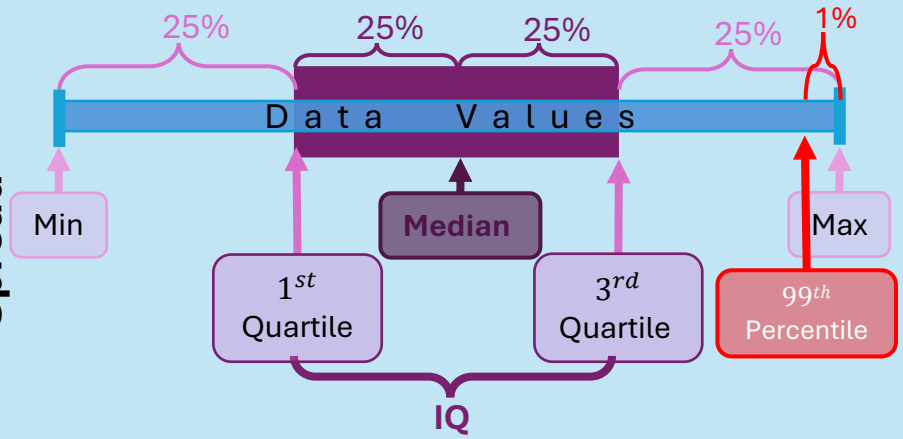
Mean (μ) = "arithmetic average value"
average value of datapoints

Sum of data values, divided by number of values

→ Pulled towards **outliers** and **tails**
→ Easy to compute and analyze

Spread

Quartiles are **data values** 25% from ends in *ordered* data

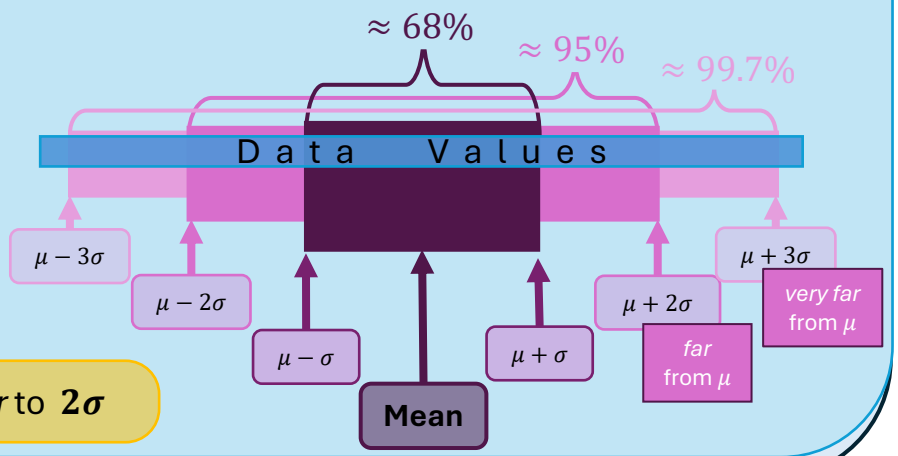


The **Inter-Quartile Range (IQ)** is the distance between the 1^{st} and 3^{rd} Quartiles.
= width of "middle 50% of data"

IQ is *similar* to 2σ

Standard Deviation (σ) is "*average* distance of data from the mean"
(formula naturally appears in many statistics computations!)

Frequently we expect data to look like this *(see Ch 7)* :

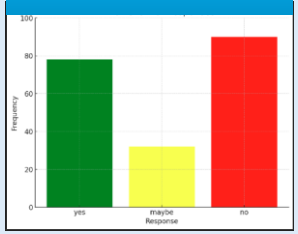


Week 2. Visualizing Data

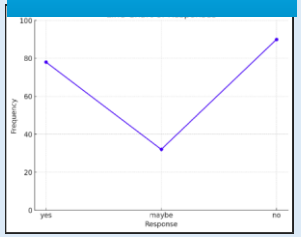
Qualitative Data:

Plot **frequencies** of each data value *(# occurrences of value)*.

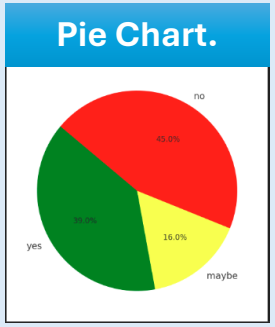
Bar Chart.



Line Chart.



Pie Chart.

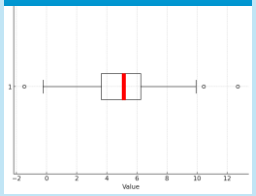


Quantitative Data:

Box Plot.

- Center line shows **median** value of data.
- Box shows **middle 50%** of data.
- Whiskers show ⊢**range**⊣ of data *(ignoring outliers)*.
- Dots show **outliers** ∘ ∘ ∘

Box Plot



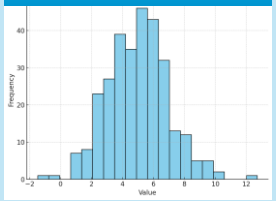
Histogram.

- Divide data into "bins"
- Count # values in each "bin"
- Make bar chart over "bins" (bars should be touching)

Tall bars ⟺ Lots of data

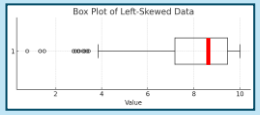
Note: choosing # bins is subtle!

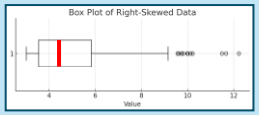
Histogram



Skew of Data

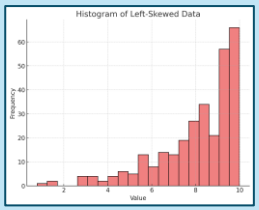
Skew data has a "tail" – going from median to mean points in direction of tail

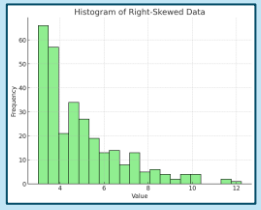




Left-Skewed *("tail" on left)*

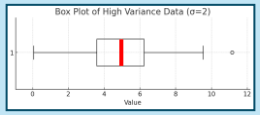
Right-Skewed *("tail" on right)*





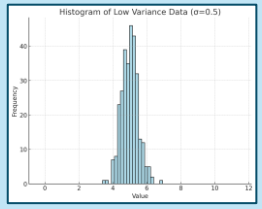
Variance of Data

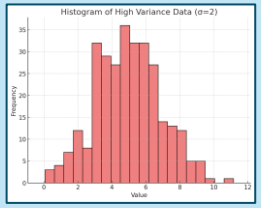
Data with high variance / standard deviation / inter-quartile range is more spread out.



Lower Variance *(gathered in)*

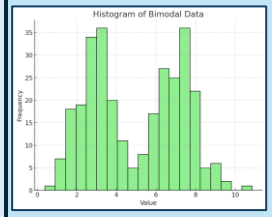
Higher Variance *(spread out)*



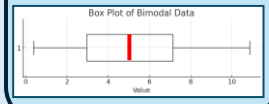


Bimodal Data

Bimodal Data has two "bumps":



Cannot be detected in box plot!



Density Estimate is an alternative to Histogram