**Variables**

* Random variable
* Numerical or categorical

**Classification 1**

* Univariate
* Multivariate

**Classification 2**

* Quantitative
  + Can be measured as ordinal data set
* Qualitative

Quantitative data are measures of values or counts and are expressed as numbers.

Quantitative data are data about numeric variables (e.g., how many; how much; or how often).

Qualitative data are measures of 'types' and may be represented by a name, symbol, or a number code.

Qualitative data are data about categorical variables (e.g., what type).

Measurements to analyse the variables:

* The size of the difference
  + Direction of reference

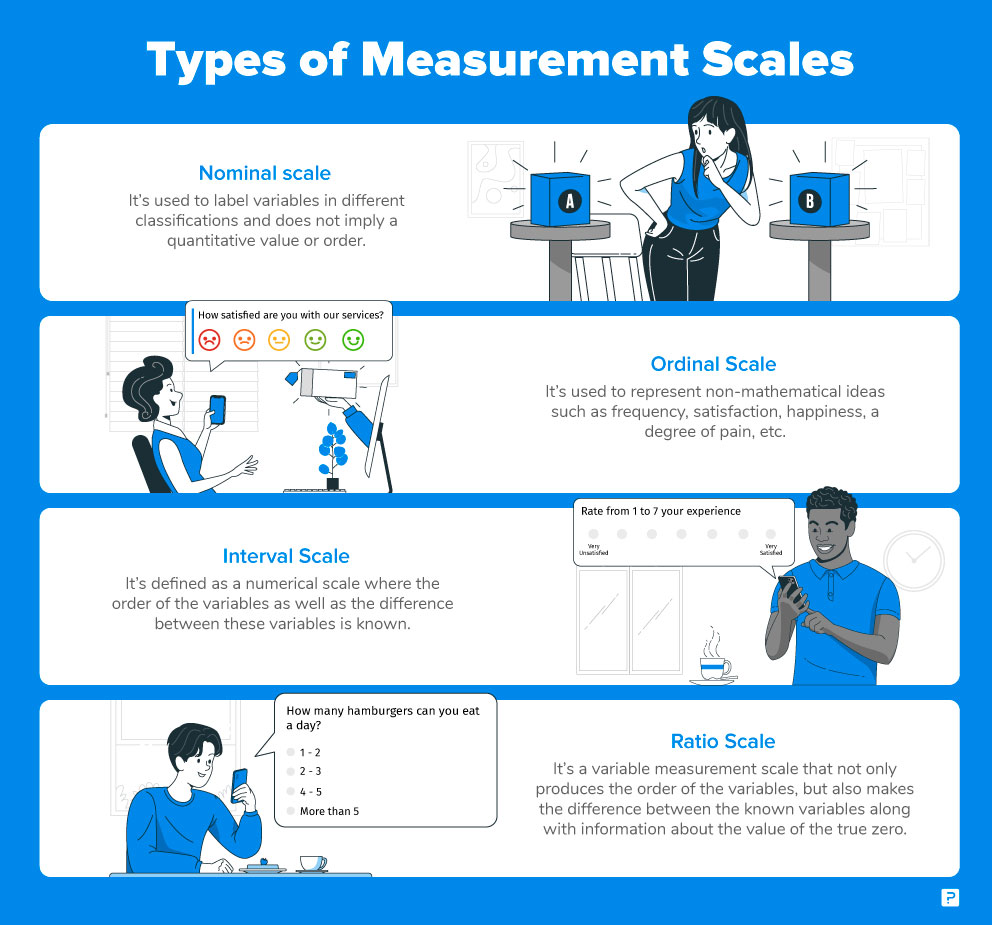
**Nominal and Ordinal scale of measurement**

[**https://www.questionpro.com/blog/nominal-vs-ordinal-scale/**](https://www.questionpro.com/blog/nominal-vs-ordinal-scale/)

[**https://www.questionpro.com/blog/nominal-ordinal-interval-ratio/**](https://www.questionpro.com/blog/nominal-ordinal-interval-ratio/)

Nominal scale is a naming scale, where variables are simply “named” or labelled, with no specific order. Ordinal scale has all its variables in a specific order, beyond just naming them. Interval scale offers labels, order, as well as a specific interval between each of its variable options. Ratio scale bears all the characteristics of an interval scale, in addition to that, it can also accommodate the value of “zero” on any of its variables.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Offers:** | **Nominal** | **Ordinal** | **Interval** | **Ratio** |
| The sequence of variables is established | – | Yes | Yes | Yes |
| Mode | Yes | Yes | Yes | Yes |
| Median | – | Yes | Yes | Yes |
| Mean | – | – | Yes | Yes |
| Difference between variables can be evaluated | – | – | Yes | Yes |
| Addition and Subtraction of variables | – | – | Yes | Yes |
| Multiplication and Division of variables | – | – | – | Yes |
| Absolute zero | – | – | – | Yes |



**Classification 3**

* Continuous
* Discrete

1. **Frequency Distributions:**
2. Percentiles and Percentile ranks
3. Grouped Frequency Distribution table

**SKEWNESS**

* If the tail points to the right , the distribution is said to be **Right skewed** , and sometimes also points towards positive numbers, hence sometimes is also called as **Positively skewed**.

(Tail is towards the right)

Chart, histogram

Description automatically generated

* If the tail points to the left , the distribution is said to be **left skewed** , and sometimes also points towards negative numbers, hence sometimes is also called as **negatively skewed**.

(Tail is towards the left)

Chart, histogram

Description automatically generated

**Comparing freq distribution**

* Grouped bar chart
* Sns.countplot() using hue
* Step type histogram
* Histtype
* Kernel density estimate plots
* Strip plots
* Box Plots
* Shows the distribution in its three quartiles.
* Outlier if value is larger than upper quartiles by 1.5 times the difference between upper quartiles and lower quartiles.
* Outlier if value is lower than lower quartiles by 1.5 times the difference between upper quartiles and lower quartiles.
* Inter Quartile range = the difference between upper quartiles and lower quartiles.
* Upper quartiles are 75%
* Middle quartiles are 50%
* Lower Quartiles are 25%
* To compare more than two distributions kde plots may not ideal ones.
* Grouped bar plots are ideal for nominal and ordinal scales
* for ratio and interval scale kde and histograms plots are better. Box plots and strip plots can be used for better readability.

**Mean Variability**

# The mean is not at the centre of the distribution. In most of the cases, the mean will not be at the centre of the distribution.

Sigma (∑) = summation

Mu (µ) = mean

Why sampling is important?

x̄ sample mean

unbiased estimator.

**Weighted Mean**

**Mean\_house\_price**

[x1,x2,x3,x4,x5]

[181762, 185139, 178842, 181405,172598]

**house\_sold**

[w1,w2,w3,w4,w5]

[625, 694, 622, 648, 341]

Weighted mean = x1w1 + x2w2 + x3w3 + x4w4 + x5w5

----------------------------------------------------

w1+w2+w3+w4+w5

**Median**

Median is considered as robust or resistant statistics as its resistant to fluctuations or changes in data.

Median is a reasonable averages for data where there exist outliers.

**mode**

data that occurs the maximum number of times.

**VARIABILITY**

Measure of Variability

* + - * Range - Difference between maximum and minimum value.
      * Having average distance checked for measure of variability.
        + Mean squared distances gives us a better understanding of variability .
* Standard Deviations - A standard deviation (or σ) is a measure of how dispersed the data is in relation to the mean. Low standard deviation means data are clustered around the mean, and high standard deviation indicates data are more spread out.

**Z SCORE**

A Z-score is a numerical measurement that describes a value's relationship to the mean of a group of values. Z-score is measured in terms of standard deviations from the mean. If a Z-score is 0, it indicates that the data point's score is identical to the mean score.

Z = (data point – mean) / Std