Statistical Analysis

- Mean
- Median
- Mode
- Stdv
- Range
- IQR
- Skewness
- Kurtosis

Statistics is the science concerned with developing and studying methods for collecting, analyzing, interpreting and presenting empirical data.

Mean, median, and mode

- Mean, median, and mode are main **measures of central tendency** in a distribution.
- Each of these measures try to **summarize** a dataset with a **single number** to represent a typical or **center data point** from the numerical data set.
- Mean good for larger sample size without outliers (symmetric dataset)
- Median is good for dataset with extreme values (skewed dataset)

Mean, median, and mode - continues

Mean: The "average" number; found by adding all data points and dividing by the number of data points.

Example: The mean of 4, 1, and 7 is (4+1+7)/3 = 12/3 = 4

Median: The middle number; found by ordering all data points and picking out the one in the middle (or if there are two middle numbers, taking the mean of those two numbers).

Example: The median of 4, 1, and 7 is 4 because when the numbers are put in order (1, 4, 7), the number 4 is in the middle.

Mean, median, and mode - continues

Mode: The most frequent number—that is, the number that occurs the highest number of times.

Example: The mode of { 4, 2, 4, 3, 2, 2} is 2 because it occurs three times, which is more than any other number.

Standard Deviation

- The **Standard Deviation** is a measure of variation or dispersion of a dataset.
- It is a number used to **tell** how measurements for a group are spread out from the average (mean), or expected value.

$$s = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2},$$

where $\{x_1, x_2, \ldots, x_N\}$ are the observed values of the sample items, $\bar{x}$ is the mean value of these observations, and N is the number of observations in the sample.

Standard Deviation - continues

The marks of a class of eight students: 2, 4, 4, 4, 5, 5, 7, 9.

These eight data points have the mean (average) of 5:

$$\mu = \frac{2+4+4+4+5+5+7+9}{8} = 5.$$

First, calculate the deviations of each data point from the mean, and square the result of each:

$$\begin{array}{ll} (2-5)^2 = (-3)^2 = 9 & (5-5)^2 = 0^2 = 0 \\ (4-5)^2 = (-1)^2 = 1 & (5-5)^2 = 0^2 = 0 \\ (4-5)^2 = (-1)^2 = 1 & (7-5)^2 = 2^2 = 4 \\ (4-5)^2 = (-1)^2 = 1 & (9-5)^2 = 4^2 = 16. \end{array}$$

The variance is the mean of these values:

$$\sigma^2 = \frac{9+1+1+1+0+0+4+16}{8} = 4.$$

and the *population* standard deviation is equal to the square root of the variance:

$$\sigma = \sqrt{4} = 2.$$

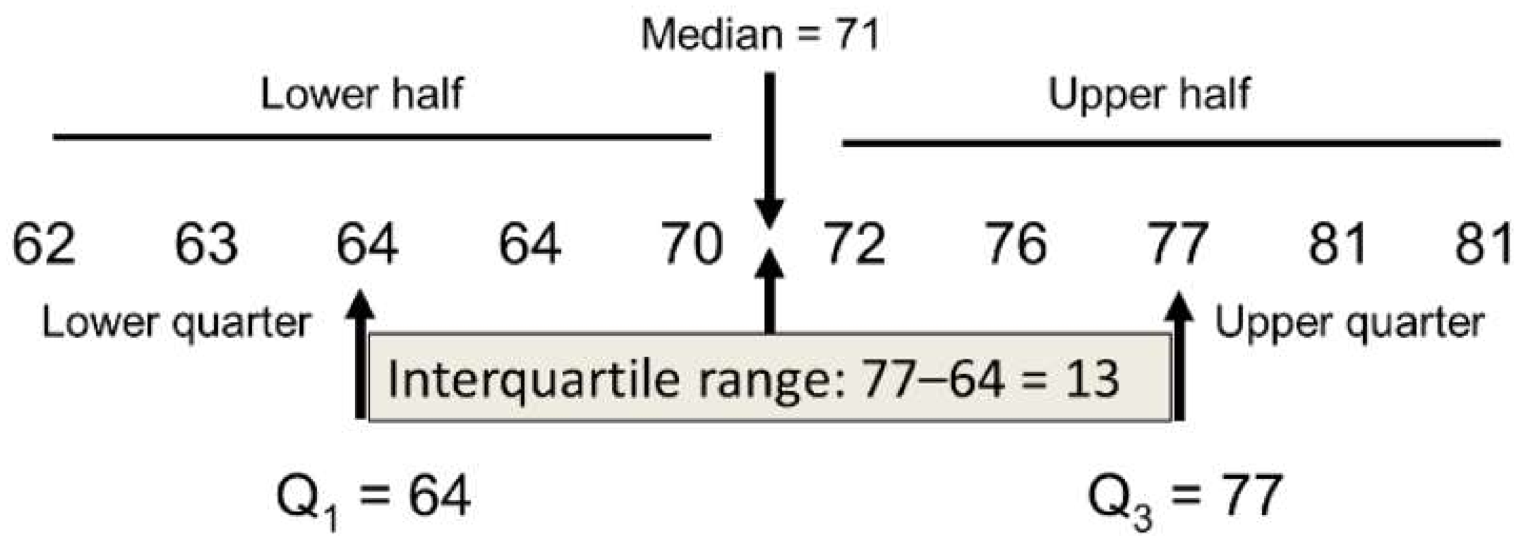
InterQuartile Range (IQR)

Interquartile Range (**IQR**)**:** It is a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles.

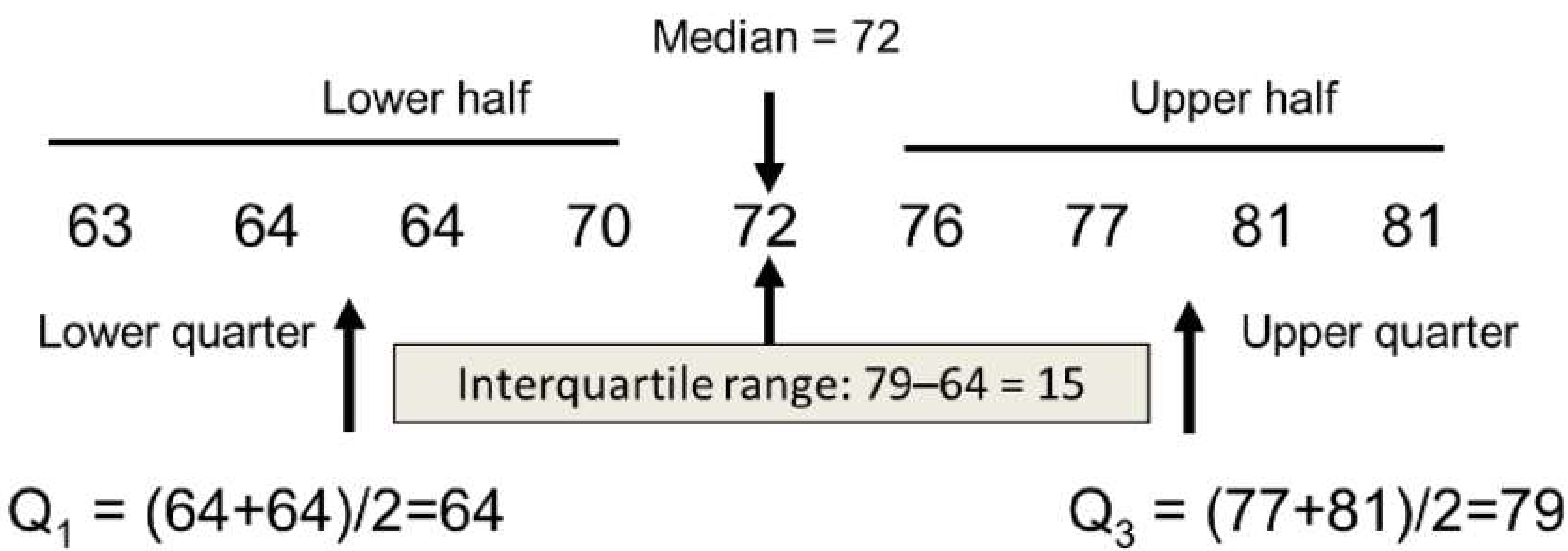
- Also called the **midspread** or **middle 50%**, or technically **H-spread**

$$\text{Interquartile Range} = Q_3 - Q_1$$

IQR- Even Sample Size

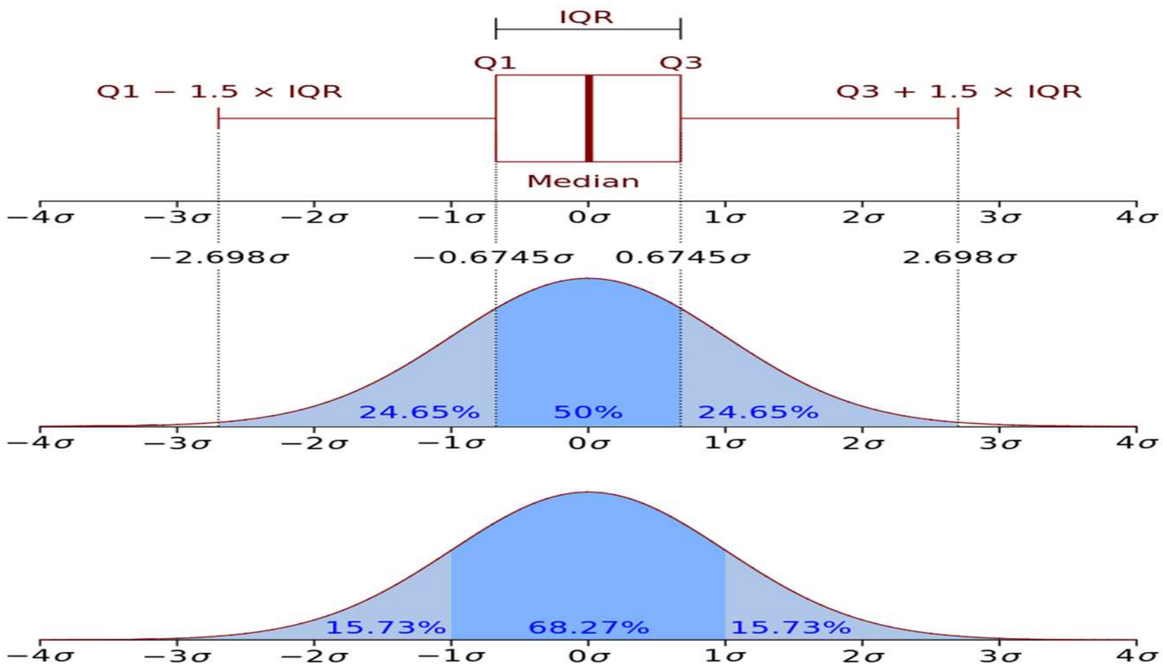


IQR- Odd Sample Size



IQR - continues

- Boxplot (with an interquartile range) and a probability density function (pdf) of a Normal N(0,σ2) Population



Summery - Statistical Analysis