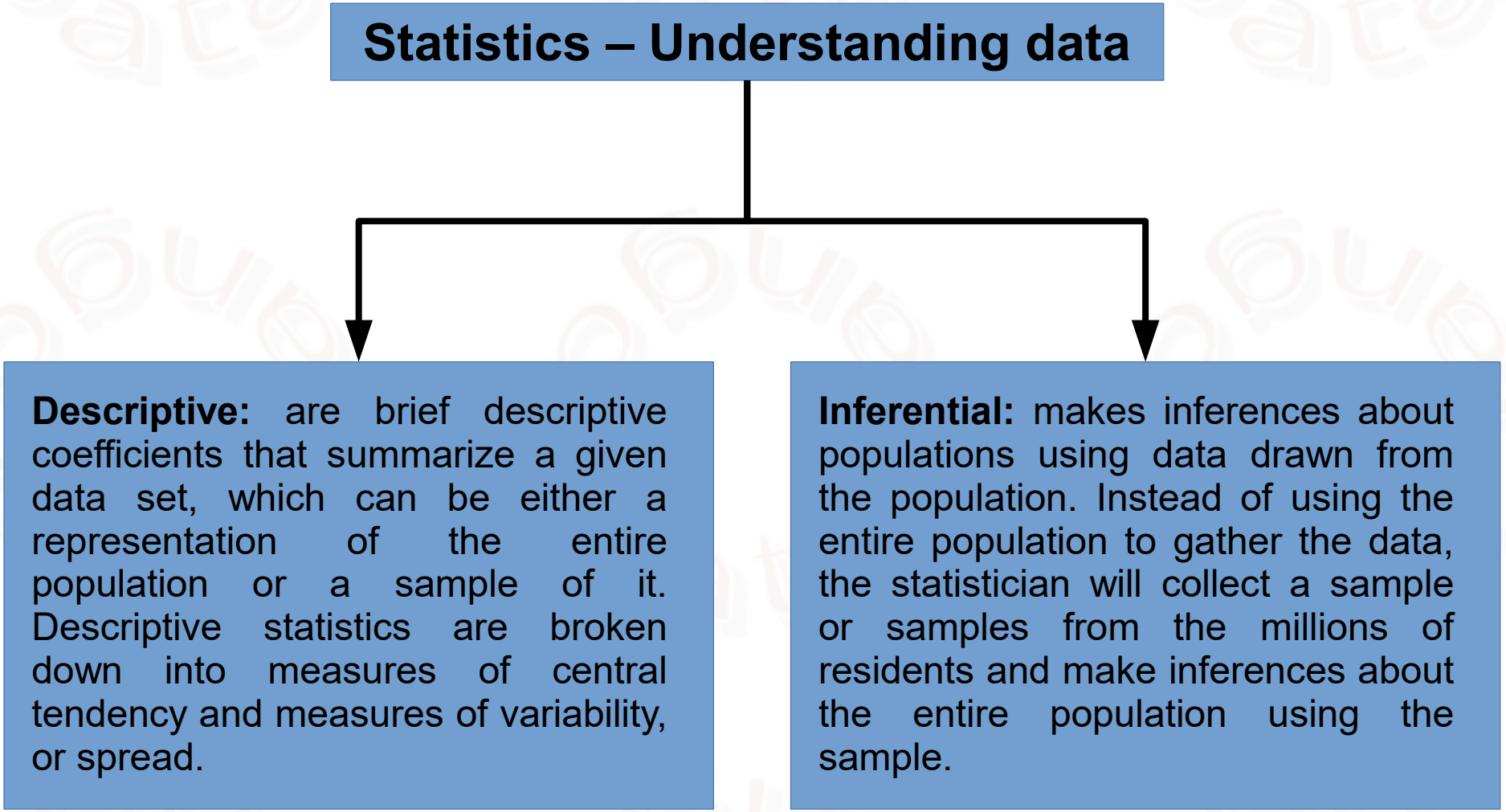
Statistics – Understanding data



Central Tendency

Mean: the mean or average that is used to derive the central tendency of the data in question. It is determined by adding all the data points in a population and then dividing the total by the number of points. The resulting number is known as the mean or the average.

Median: The median is a simple measure of central tendency. To find the median, we arrange the observations in order from smallest to largest value. If there is an odd number of observations, the median is the middle value. If there is an even number of observations, the median is the average of the two middle values.

Example : 10, 20, 3, 54, 12, 22, 70.
Sorted data : 3, 10, 12,20, 22, 54, 70
Median is '20'

- Removing outliers are not going to effect median, but will effect more on Mean.

Mode: The mode is the value that appears most often in a set of data. The mode of a discrete probability distribution is the value x at which its probability mass function takes its maximum value. In other words, it is the value that is most likely to be sampled.

Example : 3, 3, 3, 3, 3, 100
The mode is '3'

Range : how the data is spread. The difference between largest and smallest numbers.
Example: 65, 81, 73, 85, 94, 79, 67, 83, 82. Range = 94 – 65 = 29

Mid-Range: The average of largest and smallest number. 94+65/2 = 79.5

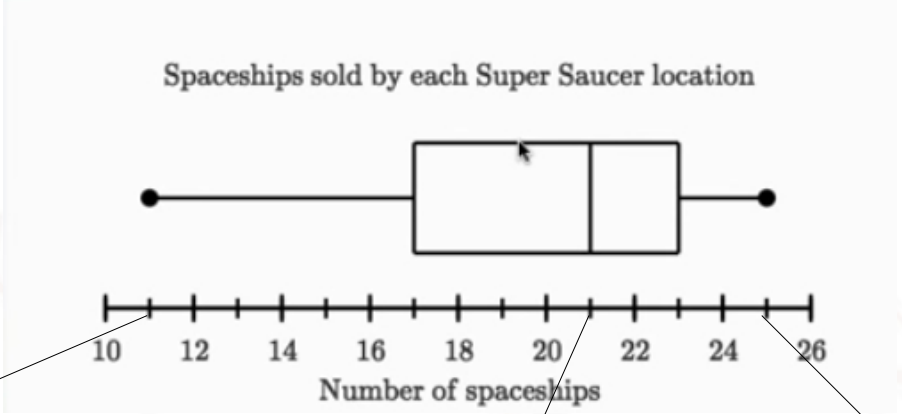
Interquartile Range: Median, middle of the first half and the middle of the second half.
A measure of spread.

Example: 4, 4, 6, 7, 10, 11, 12, 14, 15

(4+6)/2 → 5 → Quartile 1 (Q1)

10 → Median

(12+14)/2 → 13 → Quartile 3 (Q3)



Percentile: the value below which a percentage of data falls.

Example: 80% of people are shorter than you, that means you are at the 80th percentile.

English	56	75	45	71	61	64	58	80	76	61
Sorted	45	56	58	61	61	64	71	75	76	80
Percentile Rank	10	20	30	40	50	60	70	80	90	100

```
def percentileRank(scores, your_score):
    count = 0
    for score in scores:
        if score <= your_score:
            count += 1
    percentile_rank = 100.0 * count / len(scores)
    return percentile_rank
```

→ find percentile rank of 75, call *percentile_rank= percentileRank(eng, 75)* – the rank is 80

```
def getItemByRank(items, percentile_rank):
    items.sort()
    index = percentile_rank * (len(items)-1) // 100
    return items[index]
```

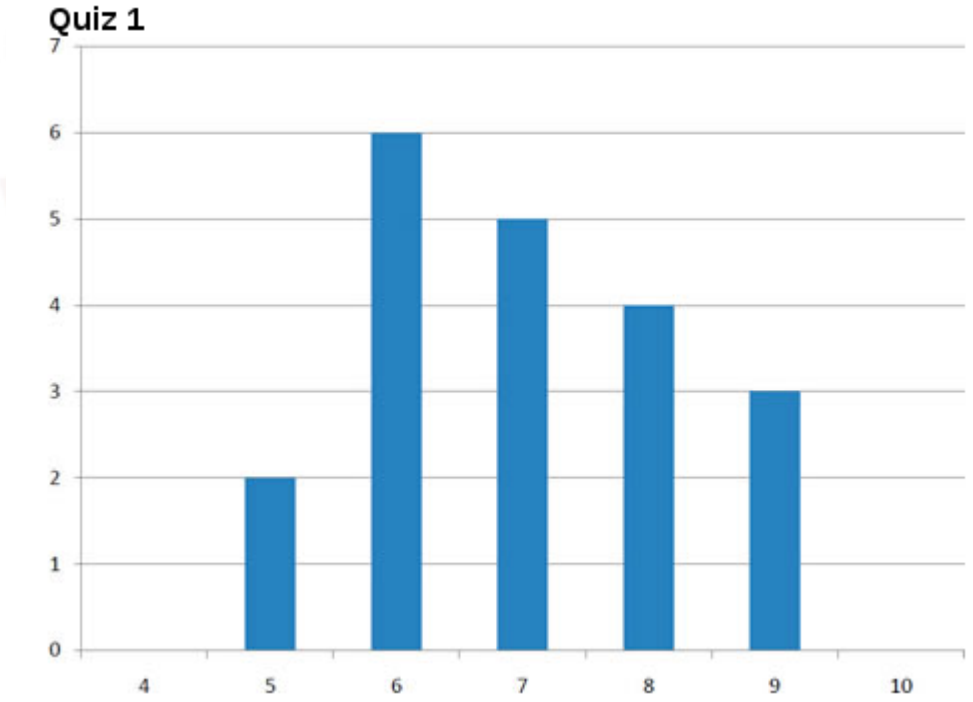
→ get the item at 60th percentile rank is *item = getItemByRank(eng, 60)* – the item is 64

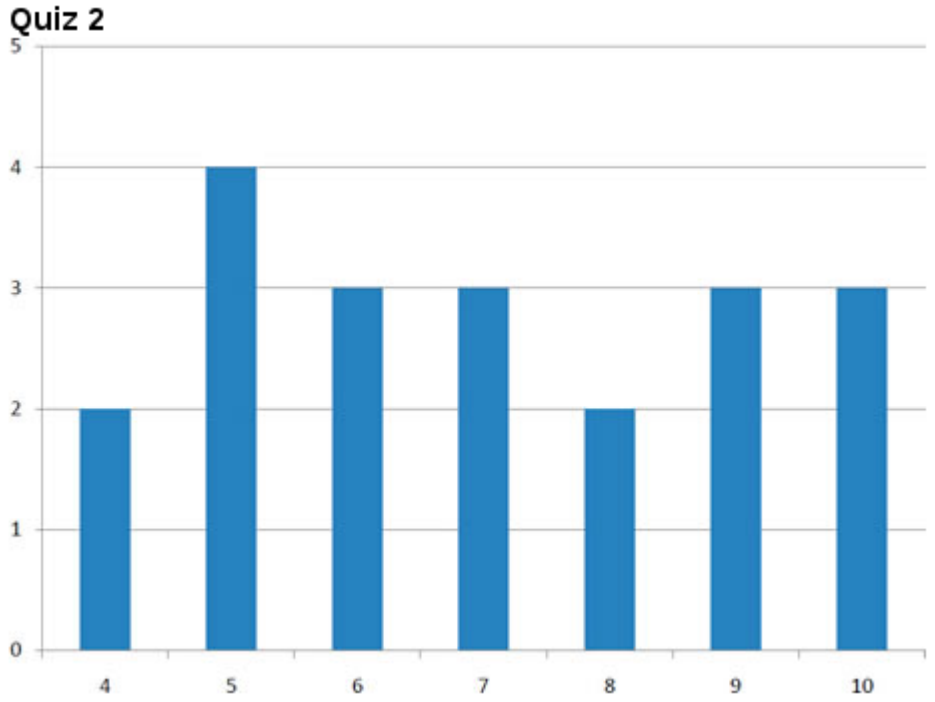
Quartile 1 = 25th percentile
Quartile 2 = 50th percentile
Quartile 3 = 75th percentile

Measures of Variability

What is Variability?

Variability refers to how "spread out" a group of scores is. To see what we mean by spread out, consider below graphs. These graphs represent the scores on two quizzes. The mean score for each quiz is 7.0. Despite the equality of means, you can see that the distributions are quite different. Specifically, the scores on Quiz 1 are more densely packed and those on Quiz 2 are more spread out. The differences among students were much greater on Quiz 2 than on Quiz 1.





x-axis : scores, y-axis : number of students

Variance: Variability can also be defined in terms of how close the scores in the distribution are to the middle of the distribution. Using the mean as the measure of the middle of the distribution, the variance is defined as the average squared difference of the scores from the mean

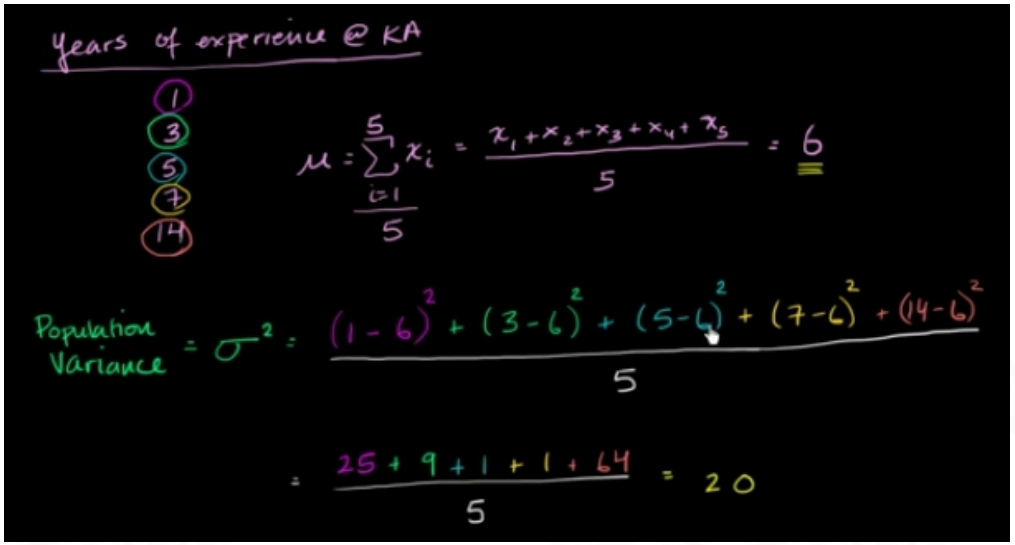
$$\sigma^2 = \frac{\sum(X-\mu)^2}{N} \quad (1)$$

$$= \frac{\sum(X^2 - 2\mu X + \mu^2)}{N} \quad (2)$$

$$= \frac{\sum X^2}{N} - \frac{2\mu\sum X}{N} + \frac{N\mu^2}{N} \quad (3)$$

$$= \frac{\sum X^2}{N} - 2\mu^2 + \mu^2 \quad (4)$$

$$= \frac{\sum X^2}{N} - \mu^2 \quad (5)$$



If the variance in a sample is used to estimate the variance in a population, then the previous formula underestimates the variance and the following formula should be used:

$$s^2 = \frac{\sum(X-M)^2}{N-1}$$

Standard Deviation: is a measure that is used to quantify the amount of variation or dispersion of a set of data values.

$$\sigma = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2}$$

Z-Score : How many standard deviations away form mean

Calculating the Standard Score (Z-Score)

$$\text{Standard Score, } z = \frac{X - \mu}{\sigma}$$

TERMS:

μ = mean (pronounced 'mu')
X = score
σ = standard deviation (pronounced 'sigma')

Co-variance: Co-variance is a measure of how much two random variables vary together. It's similar to variance, but where variance tells you how a single variable varies, co variance tells you how two variables vary together.

$$COV(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

Co-variance for the following data set:
x: 2.1, 2.5, 3.6, 4.0 (mean = 3.1)
y: 8, 10, 12, 14 (mean = 11)

Substitute the values into the formula and solve:
= (2.1-3.1)(8-11)+(2.5-3.1)(10-11)+(3.6-3.1)(12-11)+(4.0-3.1)(14-11) /(4-1)
= (-1)(-3) + (-0.6)(-1)+(.5)(1)+(0.9)(3) / 3
= 3 + 0.6 + .5 + 2.7 / 3
= 6.8/3
= 2.267

So the covariance is maximized if the two vectors are identical, 0 if they are orthogonal(uncorrelated), and negative if they point in opposite directions

Correlation coefficient: Correlation coefficients are used in statistics to measure how strong a relationship is between two variables. There are several types of correlation coefficient: Pearson's correlation or Pearson correlation is a correlation coefficient commonly used in linear regression.

$$\rho_{X,Y} = \operatorname{corr}(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sigma_X \sigma_Y}$$

Advantages of the Correlation Coefficient

The Correlation Coefficient has several advantages over covariance for determining strengths of relationships:

- Co-variance can take on practically any number while a correlation is limited: -1 to +1.
- Because of it's numerical limitations, correlation is more useful for determining how strong the relationship is between the two variables.
- Correlation does not have units. Co-variance always has units
- Correlation isn't affected by changes in the center (i.e. mean) or scale of the variables