IRON HACK

INTRODUCTION TO STATISTICS

OVERVIEW

DESCRIPTIVE STATISTICS

PAST DATA
MAKE DECISIONS
THIS IS NO ERROR OR UNCERTAINTY ASSOCIATED IN THIS PROCESS

INFERENTIAL STATISTICS

CONCLUSIONS BEYOND IMMEDIATE AVAILABLE DATA

BASED ON SAMPLE CONCLUDE FOR POPULATION

PROBABILISTIC APPROACH

PARAMETRIC ESTIMATION

CENTRAL LIMIT THEOREM

GOODNESS OF FIT TESTING

ADVANCED TOPICS

MONTECARLO SIMULATIONS
TIME SERIES
CURVE FITTING
STOCHASTIC CALCULUS

MEASURES OF DESCRIPTIVE STATISTICS

WHY DO WE NEED STATISTICAL INDICATORS OF A DATASET?

TWO TYPES: LOCATION & DISPERSION

LOCATION MEASURES CAN BE CENTRAL OR NON CENTRAL

MEASURES OF CENTRAL LOCATION

ARITHMETIC MEAN

Definition 3.3 ((Arithmetic) Mean) *The arithmetic mean (or simply mean) of a data set is given by the sum of its observations divided by the number of observations. For a sample of size n, we write*

$$x = \frac{1}{n}\sum_{i=1}^{n} x_i$$

MEDIAN

Definition 3.4 (Median) *The median is the observation located in the middle of a data set after this latter has been arranged in increasing or decreasing order.*
If the sample size n is an odd number, then the median is middle observation. If n is even, the median is the average of the two middle observations.
The median will be the number located in the

$$0.5(n+1)\text{th ordered position}$$

$$[60, 63, 65, 67, 70, 72, 75, 75, 80, 82, 84, 85]$$

$$\frac{72+75}{2} = 73.5$$

MODE

Definition 3.5 (Mode) *The mode, if it exists, is the most frequently occurring value. If many exist, then the variable is said bimodal (two) or multimodal (several). This measure fits best categorical data.*

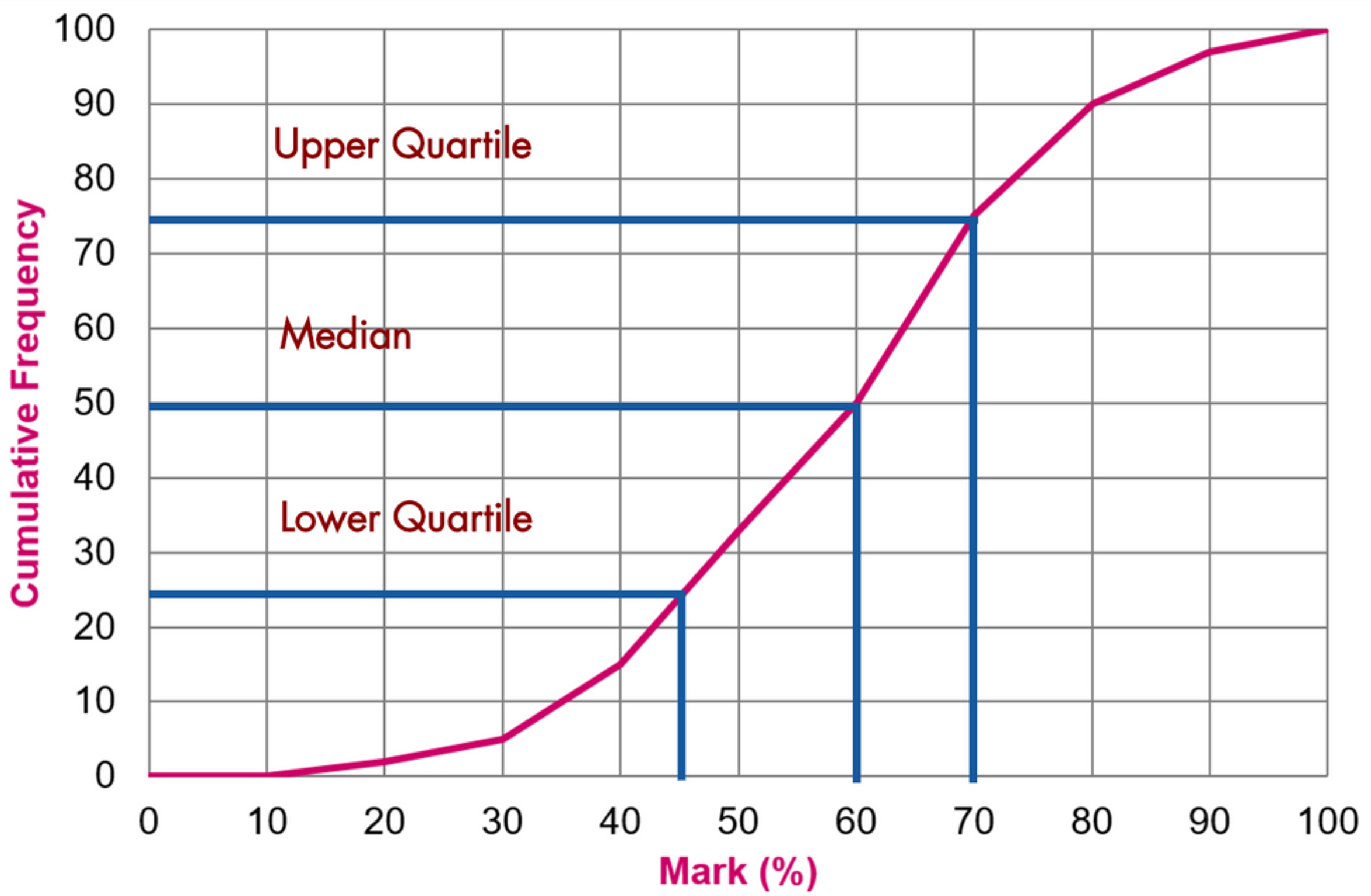
MEASURES OF NON-CENTRAL LOCATION

UPPER/LOWER QUARTILE

$$Q_1 = \text{value in the } 0.25(n+1)\text{th ordered position}$$

$$Q_2 = \text{value in the } 0.50(n+1)\text{th ordered position}$$

$$Q_3 = \text{value in the } 0.75(n+1)\text{th ordered position}$$



PERCENTILE

$$P\text{th percentile} = \text{value located in the } \frac{P}{100}(n+1)\text{th ordered position}$$

$$\text{minimum} < Q_1 < \text{median} < Q_3 < \text{maximum}$$

LET'S TEST

Exercise 3.5 Consider a sample of $n = 9$ observations that are not all equal. For the following propositions, if they are true, briefly explain why they are true. If they are false, provide a counter-example to prove they are not generally true.

a. A 10th observation is collected. If it is equal to the mean in the 9-observation sample, then the mean does not change.
b. A 10th observation is collected. If it is equal to the median in the 9-observation sample, then the median does not change.
c. A 10th observation is collected. If it is equal to the mode in the 9-observation sample, then the mode does not change.
d. A 10th and 11th observations are collected. If they are equal to the minimum and the maximum of the 9-observation sample, then the mean does not change.
e. A 10th and 11th observations are collected. If they are equal to the minimum and the maximum of the 9-observation sample, then the mode does not change.

LET'S TEST

Exercise 3.5 Consider a sample of $n = 9$ observations that are not all equal. For the following propositions, if they are true, briefly explain why they are true. If they are false, provide a counter-example to prove they are not generally true.

a. A 10th observation is collected. If it is equal to the mean in the 9-observation sample, then the mean does not change. **True**

b. A 10th observation is collected. If it is equal to the median in the 9-observation sample, then the median does not change. **True, but subtle**

c. A 10th observation is collected. If it is equal to the mode in the 9-observation sample, then the mode does not change. **True**

d. A 10th and 11th observations are collected. If they are equal to the minimum and the maximum of the 9-observation sample, then the mean does not change. **False**

e. A 10th and 11th observations are collected. If they are equal to the minimum and the maximum of the 9-observation sample, then the mode does not change. **Depends on definition of mode**

DISPERSION/VARIABILITY MEASURES

RANGE

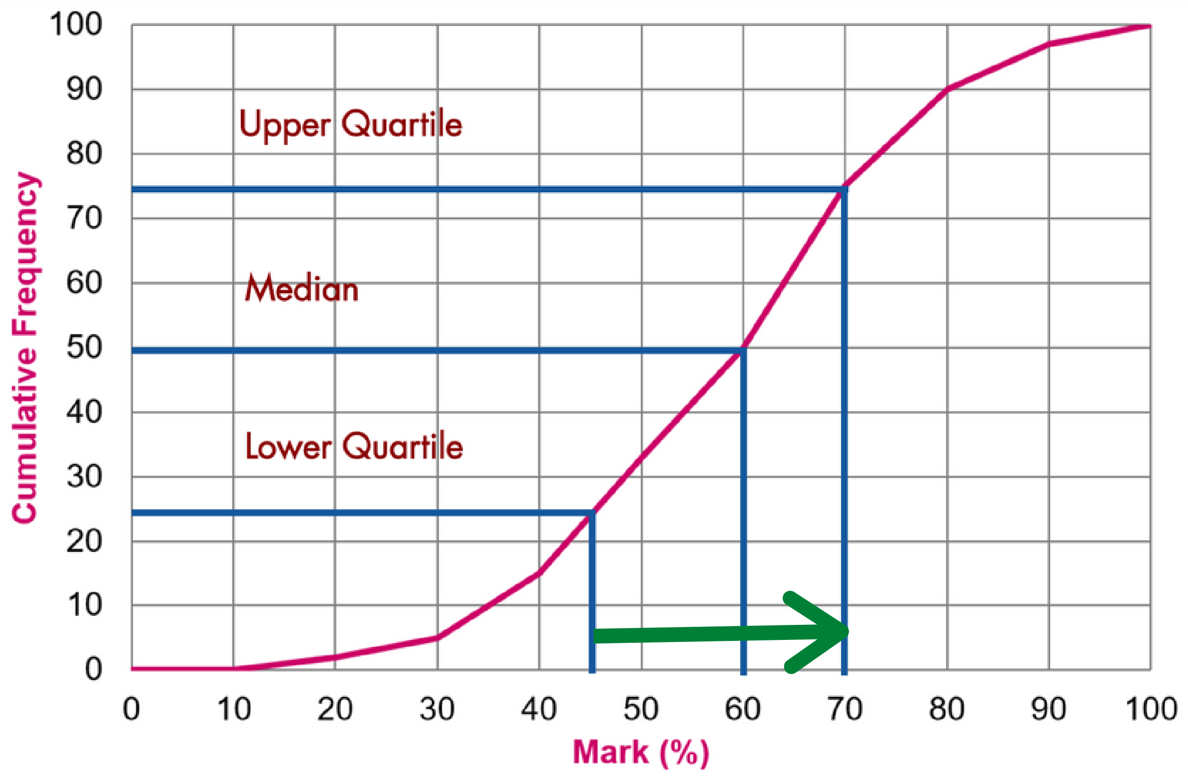
Definition 4.1 (Range) *The range of a variable is the difference between the largest and the smallest observation.*

$$[60, 63, 65, 67, 70, 72, 75, 75, 80, 82, 84, 85]$$

INTERQUARTILE RANGE

Definition 4.2 (Interquartile range) *The interquartile range (IQR) measures the spread in the middle* 50% *of the data.*
It is the difference between the value at the third quartile, Q_3, *and the observation at the first quartile,* Q_1.

$$\mathrm{IQR} = Q_3 - Q_1$$



$$[60, 63, 65, 67, 70, 72, 75, 75, 80, 82, 84, 85]$$

DISPERSION/VARIABILITY MEASURES

VARIANCE (SAMPLE AND POPULATION)

Definition 4.3 (Variance of a population) *The variance of a population, σ^2, is the sum of the squared differences of each observation with respect to the mean, divided by the population size, N.*

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

The variance of a sample of size n is based on the same differences, but the division is by $n-1$.

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - x)^2$$

ADVANCED: WHY N-1?

DISPERSION/VARIABILITY MEASURES

STANDARD DEVIATION (SAMPLE AND POPULATION)

Definition 4.4 (Standard deviation) *For a population, the standard deviation is the square root of the variance*

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

The sample's standard deviation, s. is the square root of the sample's variance.

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

ADVANCED: WHY SQRT?

COEFFICIENT OF VARIATION

Definition 4.5 (Coefficient of variation) *The coefficient of variation, CV, is a measure of relative dispersion that measures the standard deviation as a percentage of the mean (provided that the mean is positive). For a population, if* $\mu > 0$,

$$CV = \frac{\sigma}{\mu}$$

For a sample, if $x > 0$,

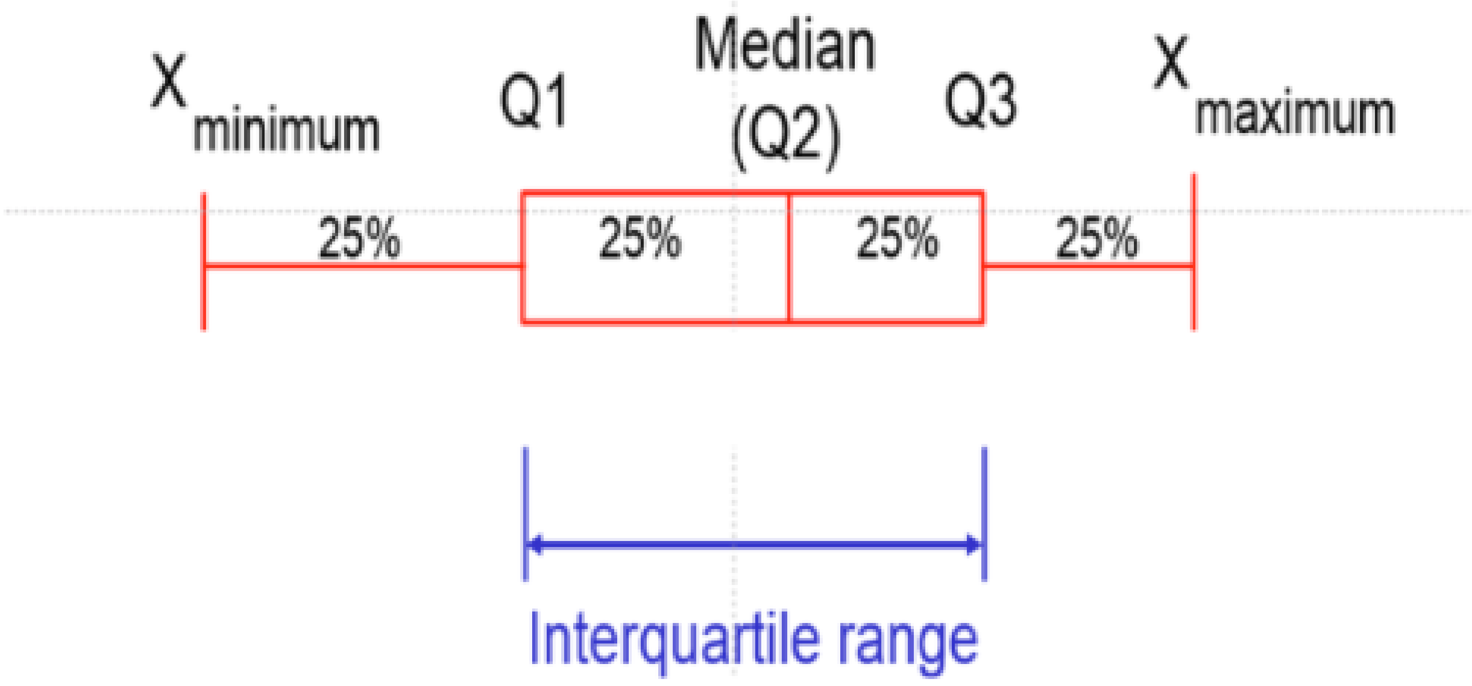
$$CV = \frac{s}{x}$$

SEE SCRIPT EXAMPLES FOR CAVEATS!

VISUAL REPRESENTATION

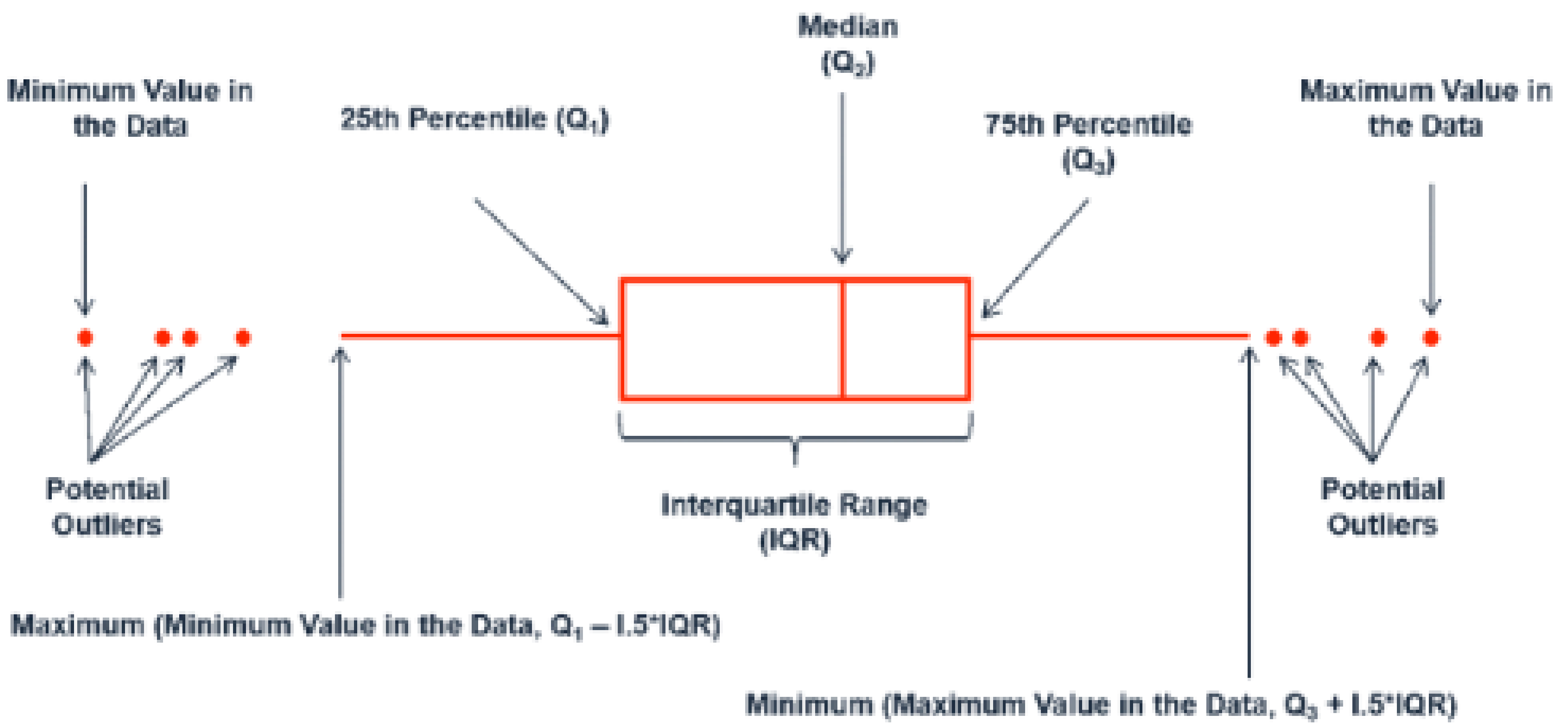
THE BOX-PLOT

This type of plot is also known as the box-and-whisker plot. There are two main alternatives for building it.



VISUAL REPRESENTATION

THE BOX-PLOT



"DEFINITION" OF OUTLIER

OUTLIER:
POINT > Q3 + 1.5*IQR

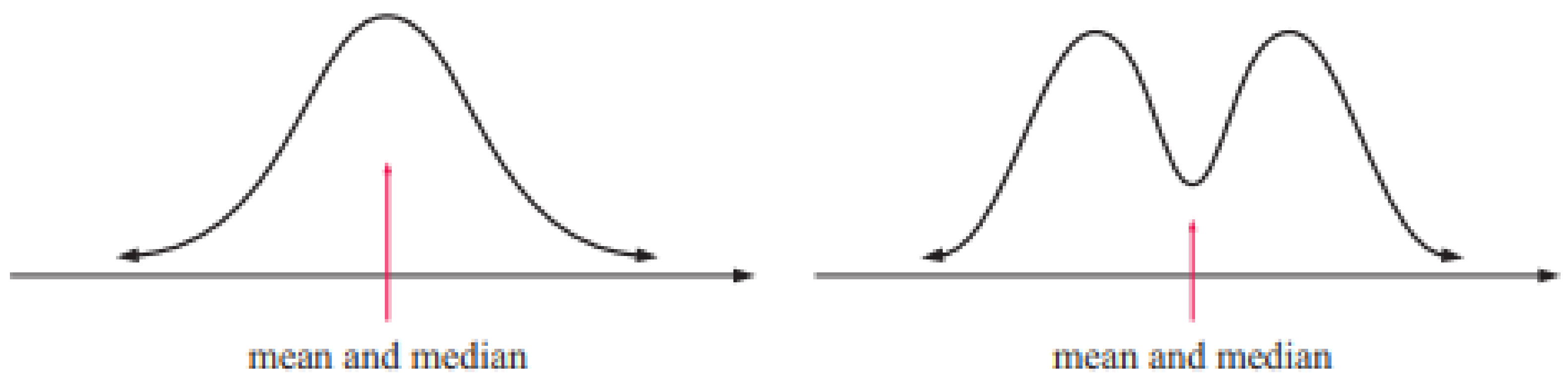
POINT < Q1 - 1.5*IQR

SHAPE OF DISTRIBUTIONS

SYMMETRY OF DISTRIBUTION OF DATA

THE RELATIONSHIP BETWEEN THE MEAN AND THE MEDIAN FOR DIFFERENT DISTRIBUTIONS

For distributions that are **symmetric** about the centre, the mean and median will be approximately equal.

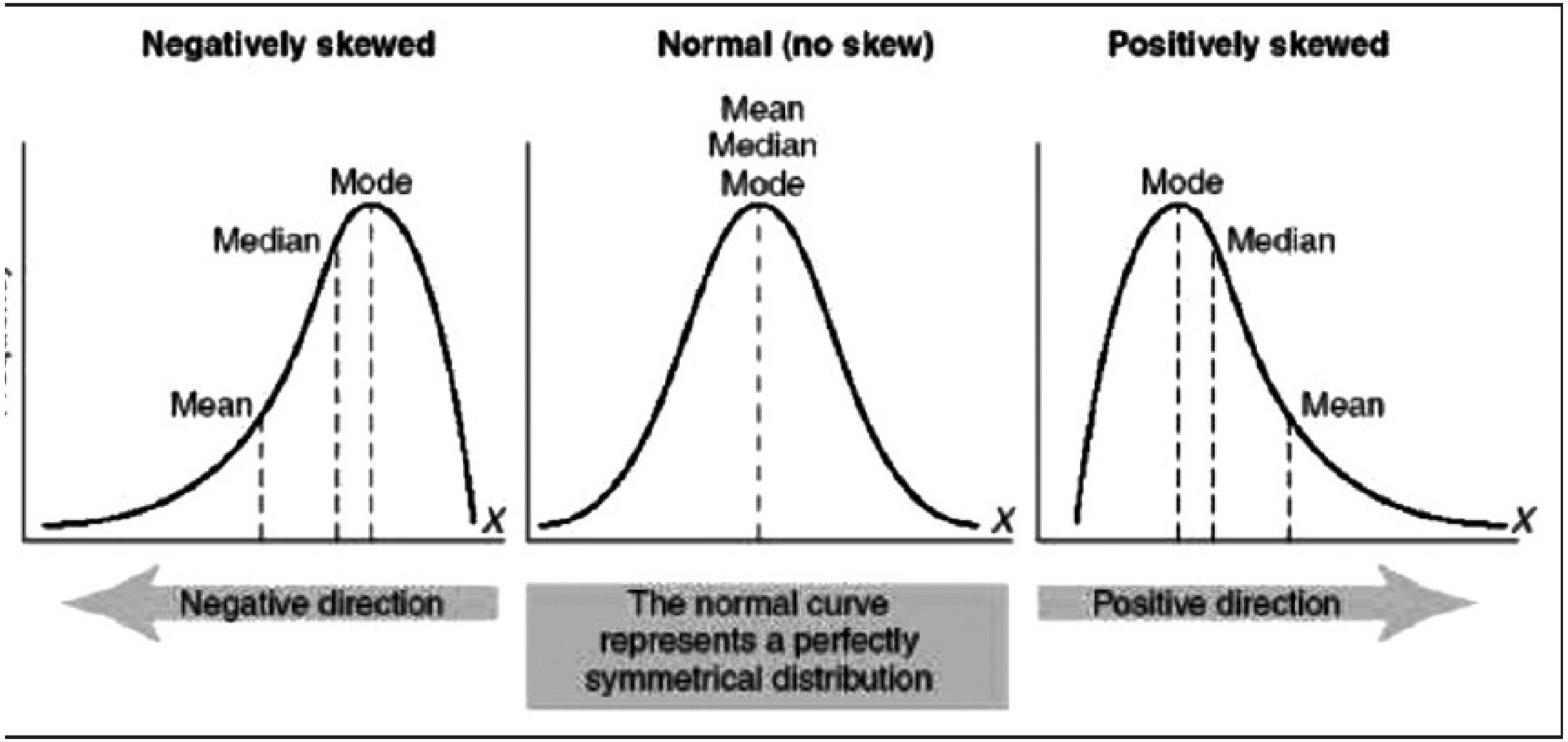


If the data set has symmetry, both the mean and the median should accurately measure the centre of the distribution.

IF SYMMETRIC: MEAN=MEDIAN

IF SYMMETRIC AND UNI-MODAL: MEAN = MEDIAN = MODE

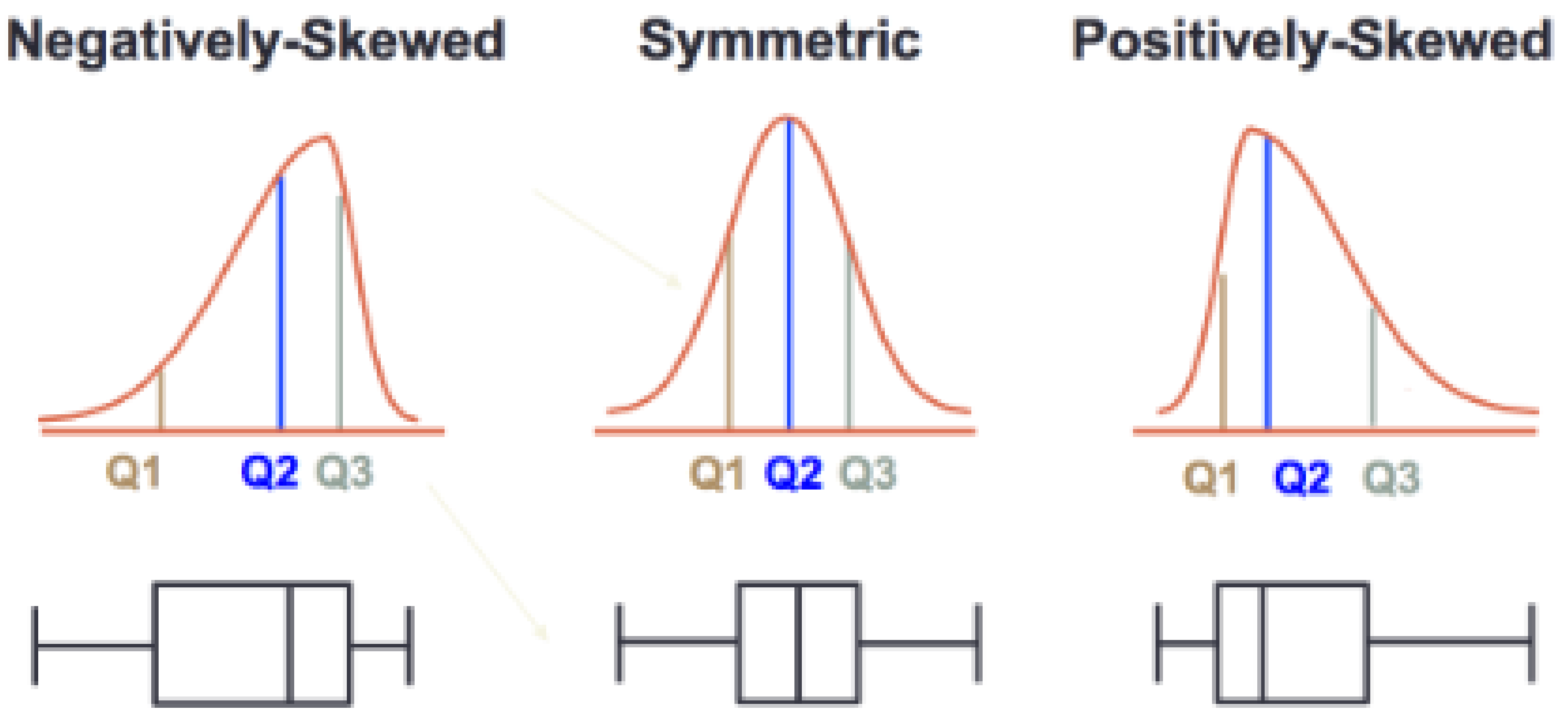
SHAPE OF DISTRIBUTIONS - SKEWDNESS AND RELATION TO INDICATORS



REMEMBER THE POWER LAWS FROM 80/20?

THE SKEWNESS OF A DISTRIBUTION OF DATA DETERMINES THE RELATIVE POSITION BETWEEN MEAN, MEDIAN AND MODE

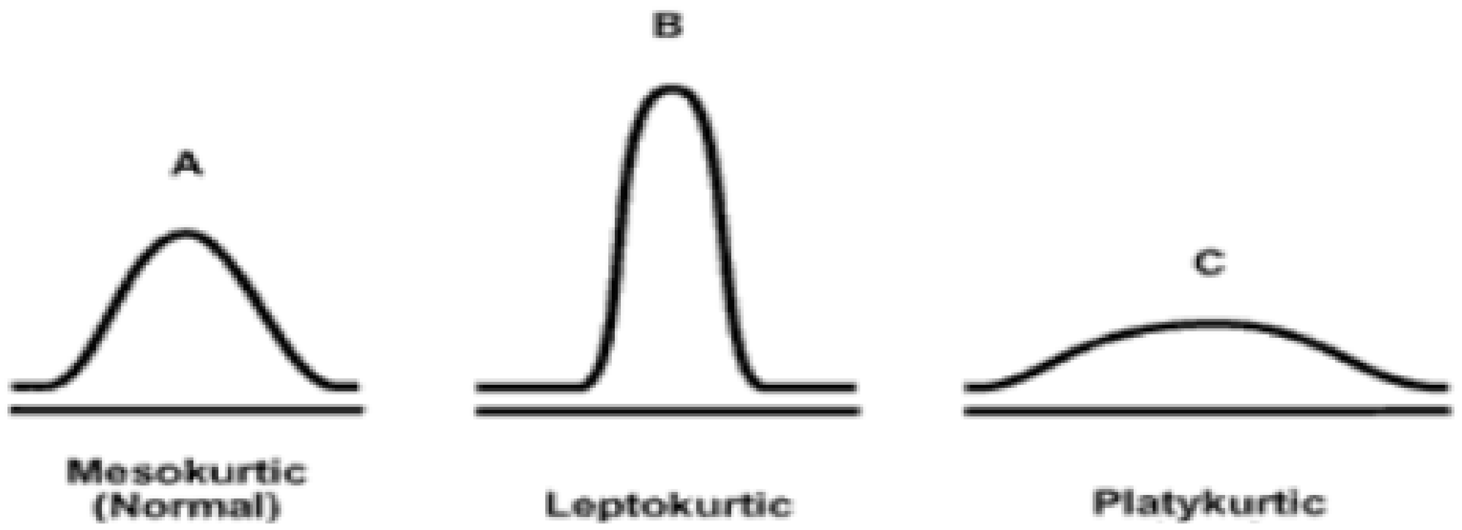
SHAPE OF DISTRIBUTIONS - RELATION TO BOX-PLOTS



KURTOSIS

Kurtosis is a measure of the combined weight of the tails relative to the rest of the distribution.

Kurtosis decreases as the tails become lighter and increases as the tails become heavier



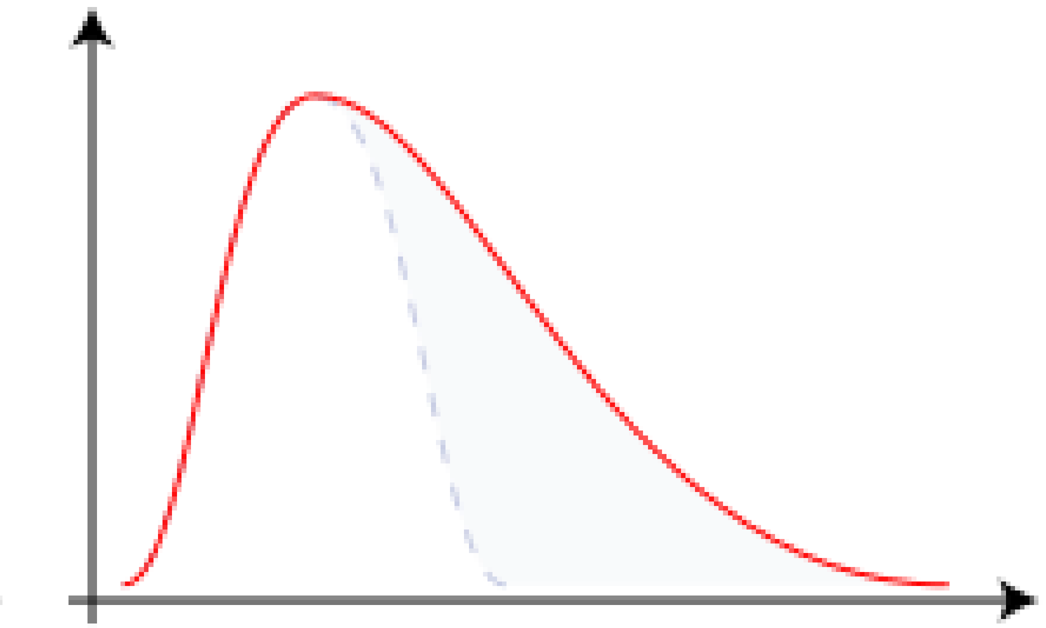
TEST TIME

Exercise 8.1 If the mean time to respond to a stimulus is much higher than the median time to respond, what can you say about the shape of the distribution of response times?

TEST TIME

Exercise 8.1 If the mean time to respond to a stimulus is much higher than the median time to respond, what can you say about the shape of the distribution of response times?

POSITIVE SKEWNESS



Positive Skew

CHEBYSHEV THEOREM

Chebyshev's Theorem

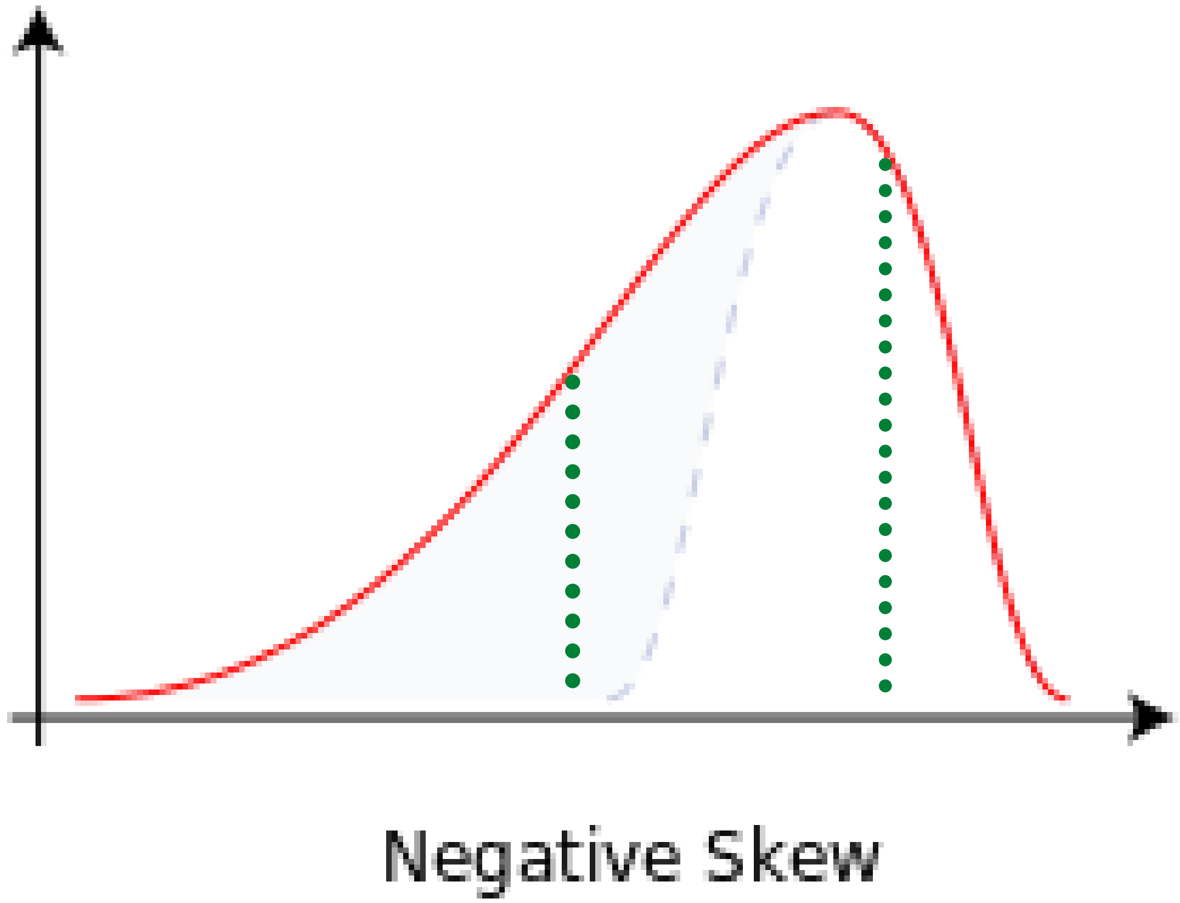
For any population with mean μ, standard deviation σ, and $k > 1$, the percent of observations that lie within the interval $[\mu \pm k\sigma]$ is

$$\textit{at least } 100[1 - (1/k^2)]\% \qquad \textbf{(2.18)}$$

where k is the number of standard deviations.

Selected Values of $k > 1$	1.5	2	2.5	3
$[1 - (1/k^2)]\%$	55.56%	75%	84%	88.89%

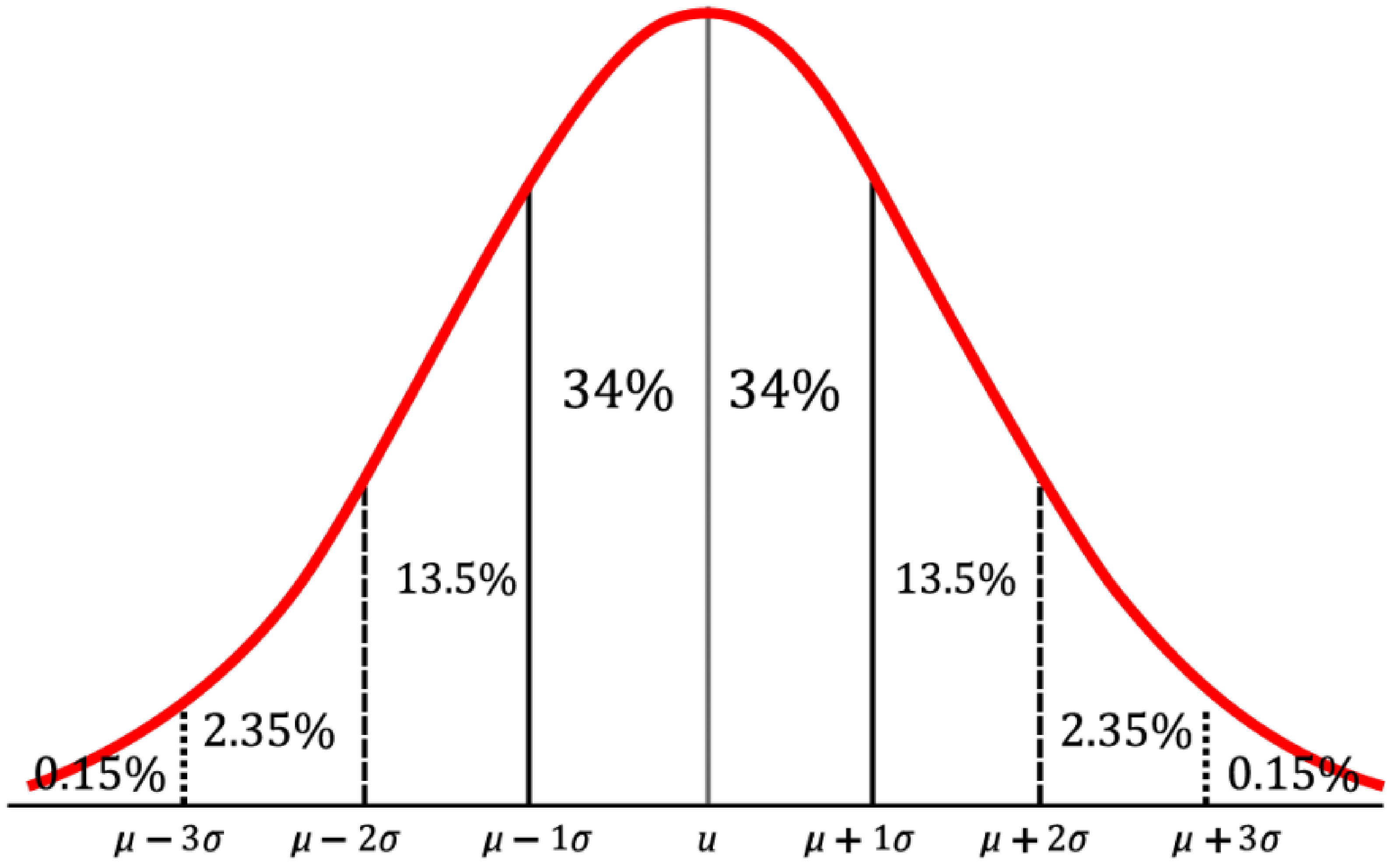
CHEBYSHEV THEOREM



Selected Values of $k > 1$	1.5	2	2.5	3
$[1 - (1/k^2)]\%$	55.56%	75%	84%	88.89%

IF THE GREEN LINES REPRESENT 1.5 STANDARD DEVIATIONS WE KNOW THAT BETWEEN THOSE LINES LIES AT LEAST 75% OF THE POINTS

IF THE CURVES ARE NORMAL (MORE ON THIS LATER)



...WE CAN DO MUCH BETTER

EMPIRICAL RULE STATES THAT FOR NORMAL DISTRIBUTIONS

68% WITHIN 1 STANDARD DEVIATION
95% WITHIN 2 STANDARD DEVIATIONS
99.7% WITHIN 3 STANDARD DEVIATIONS

NORMALIZATION

RE-SCALE THE DISTRIBUTION TO BE BETWEEN ZERO AND ONE!

EACH VALUE OF THE DATASET X WILL BE CONVERTED TO A NEW "NORMALIZED" VALUE Z

$$z = \frac{x - \min(x)}{\max(x) - \min(x)}$$

WHY?

TO COMPARE DATA IN DIFFERENT SCALES

E.G. DISTANCE AND SIMILARITY BASED METRICS BETWEEN

	Height	Salary	Age
Eenie	1.72	27000.0	29.0
Meeni	1.81	25000.0	32.0
Miny	1.79	32000.0	41.0

STANDARDIZATION

THE Z-SCORE IS A WAY TO STANDARDIZE ALL YOUR DATA IN A WAY THAT TELLS YOU HOW MANY STANDARD DEVIATIONS EACH POINT IS FROM THE MEAN

$$Z = \frac{x - \mu}{\sigma}$$

WHY?

BECAUSE AN "EXTREME" NORMALIZED VALUE MAY NOT BE THAT EXTREME AFTER ALL

COVARIANCE

COVARIANCE IS A WAY TO DESCRIBE THE LINEAR RELATIONSHIP BETWEEN TWO VARIABLES

$$Cov(x, y) = \sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

POSITIVE COVARIANCE -> VARIABLES CHANGE IN THE SAME DIRECTION: ONE GOES UP, THE OTHER GOES UP. E.G. EDUCATION AND SALARY

NEGATIVE COVARIANCE -> ANTI-CORRELATION BETWEEN VARIABLES. ONE GOES UP, THE OTHER GOES DOWN. E.G. EDUCATION AND HOMELESSNESS

THE COVARIANCE DESCRIBES ONLY THE DIRECTION THE LINEAR CORRELATION BETWEEN BOTH VARIABLES AND NOT THE STRENGTH

WHY? THINK OF THE UNITS! ARE THEY RELATIVE OR ABSOLUTE?

THEREFORE COVARIANCE SHOULD ONLY BE USED IN SITUATIONS WHERE THE UNITS WILL BE MADE MEANINGFUL FURTHER DOWN THE LINE. FOR A DIRECTLY INTERPRETABLE RESULT WE SHOULD USE CORRELATION

COVARIANCE

EXAMPLE FROM TEXTBOOK

Example 2.19 Facebook Posts and Interactions (Covariance and Correlation Coefficient)

RELEVANT Magazine (a culture magazine) keeps in touch and informs their readers by posting updates through various social networks. These updates take up a large part of both the marketing and editorial teams' time. Because these updates take so much time, marketing is interested in knowing whether reducing posts (updates) on Facebook (a specific site) will also lessen their fan interaction; if not, both departments may pursue using their time in more productive ways. The weekly number of posts (updates) and fan interactions for Facebook during a 9-week period are recorded in Table 2.10. Compute the covariance and correlation between Facebook posts (site updates) and fan interactions. The data are stored in the data file **RELEVANT Magazine**.

COVARIANCE

EXAMPLE FROM TEXTBOOK

Example 2.19 Facebook Posts and Interactions (Covariance and Correlation Coefficient)

RELEVANT Magazine (a culture magazine) keeps in touch and informs their readers by posting updates through various social networks. These updates take up a large part of both the marketing and editorial teams' time. Because these updates take so much time, marketing is interested in knowing whether reducing posts (updates) on Facebook (a specific site) will also lessen their fan interaction; if not, both departments may pursue using their time in more productive ways. The weekly number of posts (updates) and fan interactions for Facebook during a 9-week period are recorded in Table 2.10. Compute the covariance and correlation between Facebook posts (site updates) and fan interactions. The data are stored in the data file **RELEVANT Magazine**.

$$Cov(x, y) = \sigma_{xy} = \frac{\sum_{i=1}^{N} (x_i - \mu_x)(y_i - \mu_y)}{N}$$

Table 2.10 Facebook Posts (site updates) and Fan Interactions

Facebook posts (updates), x	16	31	27	23	15	17	17	18	14
Fan interactions, y	165	314	280	195	137	286	199	128	462

Solution The computation of covariance and correlation between Facebook posts (site updates) and fan interactions are illustrated in Table 2.11. The mean and the variance in the number of Facebook posts are found to be approximately

$$\bar{x} = 19.8 \quad \text{and} \quad s_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1} = 34.694$$

and the mean and the variance in the number of fan interactions are found to be approximately

$$\bar{y} = 240.7 \quad \text{and} \quad s_y^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n-1} = 11{,}369.5$$

Table 2.11 Facebook Posts and Fan Interactions (Covariance and Correlation)

x	y	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
16	165	−3.8	14.44	−75.7	5,730.49	287.66
31	314	11.2	125.44	73.3	5,372.89	820.96
27	280	7.2	51.84	39.3	1,544.49	282.96
23	195	3.2	10.24	−45.7	2,088.49	−146.24
15	137	−4.8	23.04	−103.7	10,753.69	497.76
17	286	−2.8	7.84	45.3	2,052.09	−126.84
17	199	−2.8	7.84	−41.7	1,738.89	116.76
18	128	−1.8	3.24	−112.7	12,701.29	202.86
14	462	−5.8	33.64	221.3	48,973.69	−1,283.54
$\bar{x} = 19.8$	$\bar{y} = 240.7$					$\sum = 652.34$

COVARIANCE

APPLYING THE EQUATION ABOVE (FOR SAMPLE)

$$Cov(x, y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{652.34}{8} = 81.542$$

Table 2.10 Facebook Posts (site updates) and Fan Interactions

Facebook posts (updates), x	16	31	27	23	15	17	17	18	14
Fan interactions, y	165	314	280	195	137	286	199	128	462

Solution The computation of covariance and correlation between Facebook posts (site updates) and fan interactions are illustrated in Table 2.11. The mean and the variance in the number of Facebook posts are found to be approximately

$$\bar{x} = 19.8 \quad \text{and} \quad s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1} = 34.694$$

and the mean and the variance in the number of fan interactions are found to be approximately

$$\bar{y} = 240.7 \quad \text{and} \quad s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n - 1} = 11{,}369.5$$

Table 2.11 Facebook Posts and Fan Interactions (Covariance and Correlation)

x	y	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
16	165	–3.8	14.44	–75.7	5,730.49	287.66
31	314	11.2	125.44	73.3	5,372.89	820.96
27	280	7.2	51.84	39.3	1,544.49	282.96
23	195	3.2	10.24	–45.7	2,088.49	–146.24
15	137	–4.8	23.04	–103.7	10,753.69	497.76
17	286	–2.8	7.84	45.3	2,052.09	–126.84
17	199	–2.8	7.84	–41.7	1,738.89	116.76
18	128	–1.8	3.24	–112.7	12,701.29	202.86
14	462	–5.8	33.64	221.3	48,973.69	–1,283.54
$\bar{x} = 19.8$	$\bar{y} = 240.7$					$\Sigma = 652.34$

PEARSON CORRELATION

THE CORRELATION PROVIDES THE DIRECTION AND STRENGTH OF THE LINEAR RELATIONSHIP BETWEEN TWO VARIABLES

$$\rho = \frac{Cov(x, y)}{\sigma_x \sigma_y}$$

UNLIKE THE COVARIANCE, THE CORRELATION IS A RELATIVE MEASURE AND THEREFORE CAN BE USED AS COMPARISON BETWEEN DIFFERENT MEASURES/DATASETS

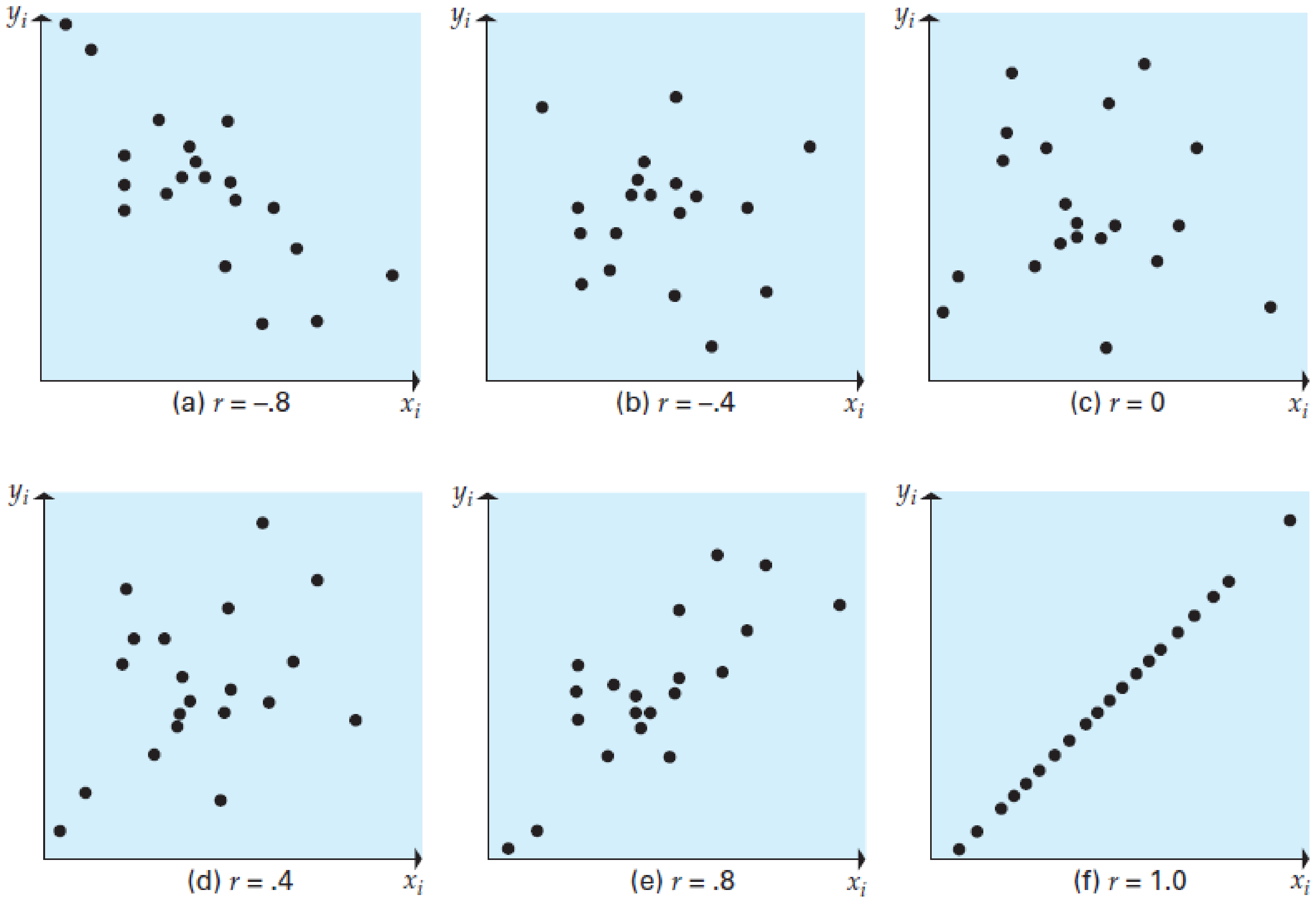
CORRELATION GIVES US A STANDARDIZED MEASURE OF THE LINEAR RELATIONSHIP BETWEEN THE TWO VARIABLES

+1 INDICATES A PERFECT LINEAR CORRELATION

-1 INDICATES A PERCECT LINEAR ANTI-CORRELATION

PEARSON CORRELATION

THE CORRELATION PROVIDES THE DIRECTION AND STRENGTH OF THE LINEAR RELATIONSHIP BETWEEN TWO VARIABLES



WRAPPING IT UP

$$Cov(x, y) = s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{652.34}{8} = 81.542$$

$$r = \frac{Cov(x, y)}{s_x s_y} = \frac{81.542}{\sqrt{34.694}\sqrt{11{,}369.5}} = 0.1298$$

DATA DOES NOT SUPPORT STRON LINEAR RELATIONSHIP BETWEEN POSTS AND FAN INTERACTION

Table 2.10 Facebook Posts (site updates) and Fan Interactions

Facebook posts (updates), x	16	31	27	23	15	17	17	18	14
Fan interactions, y	165	314	280	195	137	286	199	128	462

Solution The computation of covariance and correlation between Facebook posts (site updates) and fan interactions are illustrated in Table 2.11. The mean and the variance in the number of Facebook posts are found to be approximately

$$\bar{x} = 19.8 \quad \text{and} \quad s_x^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1} = 34.694$$

and the mean and the variance in the number of fan interactions are found to be approximately

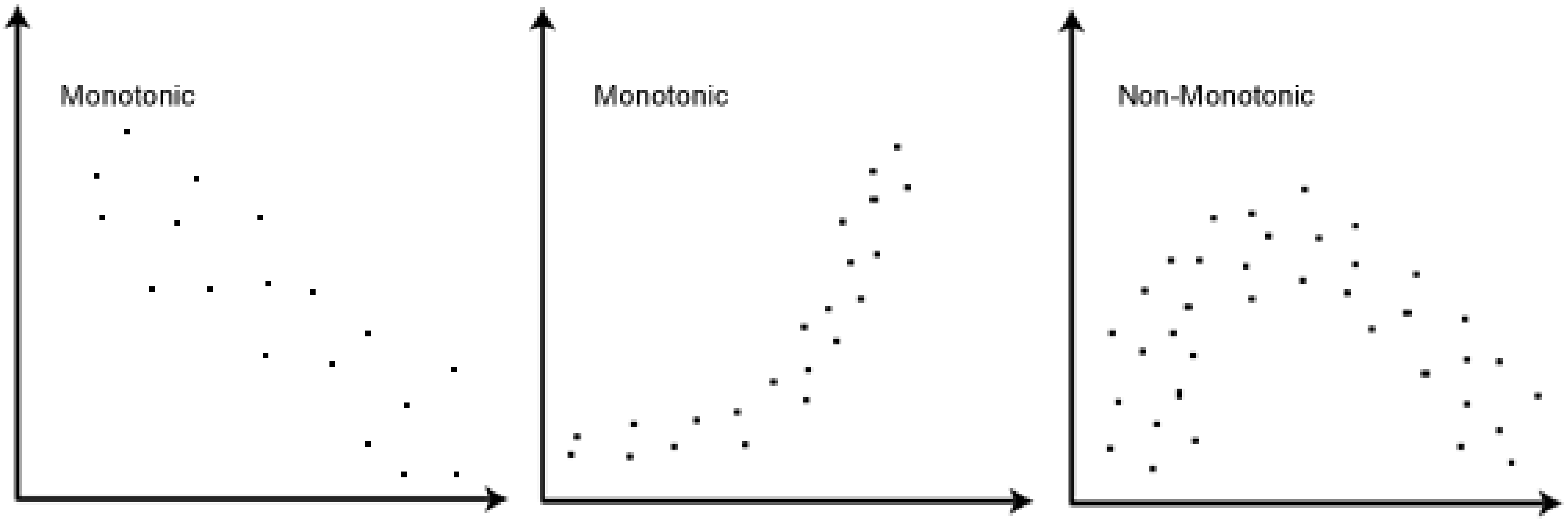
$$\bar{y} = 240.7 \quad \text{and} \quad s_y^2 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n - 1} = 11{,}369.5$$

Table 2.11 Facebook Posts and Fan Interactions (Covariance and Correlation)

x	y	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$
16	165	–3.8	14.44	–75.7	5,730.49	287.66
31	314	11.2	125.44	73.3	5,372.89	820.96
27	280	7.2	51.84	39.3	1,544.49	282.96
23	195	3.2	10.24	–45.7	2,088.49	–146.24
15	137	–4.8	23.04	–103.7	10,753.69	497.76
17	286	–2.8	7.84	45.3	2,052.09	–126.84
17	199	–2.8	7.84	–41.7	1,738.89	116.76
18	128	–1.8	3.24	–112.7	12,701.29	202.86
14	462	–5.8	33.64	221.3	48,973.69	–1,283.54
$\bar{x} = 19.8$	$\bar{y} = 240.7$					$\Sigma = 652.34$

SPEARMAN CORRELATION

THE SPEARMAN CORRELATION DOESN'T LOOK FOR A LINEAR RELATIONSHIP BUT RATHER A MONOTONIC RELATIONSHIP (IN THE SAME DIRECTION). IT DOES SO BY COMPARING THE RANK OF THE POINTS



$$\rho_{\mathrm{rg}_X,\mathrm{rg}_Y} = \frac{\mathrm{cov}(\mathrm{rg}_X, \mathrm{rg}_Y)}{\sigma_{\mathrm{rg}_X}\sigma_{\mathrm{rg}_Y}}$$

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

IN SPEARMAN THE RANK OF EACH POINT IS WHAT MATTERS RATHER THAN THE ABSOLUTE VALUE

SPEARMAN CORRELATION

LET'S CALCULATE THE SPEARMAN CORRELATION BETWEEN THE GRADES OF A CLASS IN MATHS AND ENGLISH

	Marks									
English	56	75	45	71	61	64	58	80	76	61
Maths	66	70	40	60	65	56	59	77	67	63

SPEARMAN CORRELATION

START BY RANKING THEM WITHIN THE RESPECTIVE COLUMN

English (mark)	Maths (mark)	Rank (English)	Rank (maths)
56	66	9	4
75	70	3	2
45	40	10	10
71	60	4	7
61	65	6.5	5
64	56	5	9
58	59	8	8
80	77	1	1
76	67	2	3
61	63	6.5	6

SPEARMAN CORRELATION

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

CALCULATE THE DIFFERENCE OF THE RANK FOR EACH STUDENT

English (mark)	Maths (mark)	Rank (English)	Rank (maths)	d	d^2
56	66	9	4	5	25
75	70	3	2	1	1
45	40	10	10	0	0
71	60	4	7	3	9
62	65	6	5	1	1
64	56	5	9	4	16
58	59	8	8	0	0
80	77	1	1	0	0
76	67	2	3	1	1
61	63	7	6	1	1

$$\sum d_i^2 = 25+1+9+1+16+1+1 = 54$$

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

$$\rho = 1 - \frac{6 \times 54}{10(10^2-1)}$$

$$\rho = 1 - \frac{324}{990}$$

$$\rho = 1 - 0.33$$

$$\rho = 0.67$$

KENDALL TAU CORRELATION

SIMILARLY TO SPEARMAN'S CORRELATION KENDAL TAU'S DOESN'T LOOK FOR A LINEAR RELATIONSHIP BUT RATHER A MONOTONIC RELATIONSHIP (IN THE SAME DIRECTION). IT DOES SO BY COMPARING THE RANK OF THE POINTS

$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\binom{n}{2}}.$$ [3]

Where $\binom{n}{2} = \frac{n(n-1)}{2}$ is the binomial coefficient for the number of ways to choose two items from n items.

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be a set of observations of the joint random variables X and Y respectively, such that all the values of (x_i) and (y_i) are unique. Any pair of observations (x_i, y_i) and (x_j, y_j), where $i < j$, are said to be *concordant* if the ranks for both elements (more precisely, the sort order by x and by y) agree: that is, if both $x_i > x_j$ and $y_i > y_j$; or if both $x_i < x_j$ and $y_i < y_j$. They are said to be *discordant*, if $x_i > x_j$ and $y_i < y_j$; or if $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant.

KENDALL TAU CORRELATION

SIMILARLY TO SPEARMAN'S CORRELATION KENDAL TAU'S DOESN'T LOOK FOR A LINEAR RELATIONSHIP BUT RATHER A MONOTONIC RELATIONSHIP (IN THE SAME DIRECTION). IT DOES SO BY COMPARING THE RANK OF THE POINTS

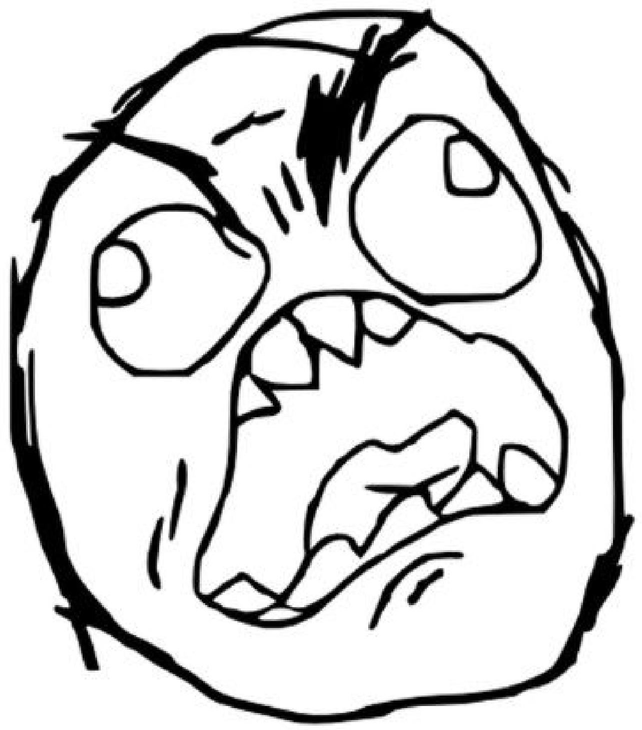
$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\binom{n}{2}}.$$ [3]

Where $\binom{n}{2} = \frac{n(n-1)}{2}$ is the binomial coefficient for the number of ways to choose two items from n items.

Let $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ be a set of observations of the joint random variables *X* and *Y* respectively, such that all the values of (and (y_i) are unique. Any pair of observations (x_i, y_i) and (x_j, y_j), where $i < j$, are said to be *concordant* if the ranks for both elements (more precisely, the sort order by *x* and by *y*) agree: that is, if both $x_i > x_j$ and $y_i > y_j$; or if both $x_i < x_j$ and $y_i < y_j$. They are said to be *discordant*, if $x_i > x_j$ and $y_i < y_j$; or if $x_i < x_j$ and $y_i > y_j$. If $x_i = x_j$ or $y_i = y_j$, the pair is neither concordant nor discordant.

KENDALL TAU CORRELATION

SIMILARLY TO SPEARMAN'S CORRELATION KENDAL TAU'S DOESN'T LOOK FOR A LINEAR RELATIONSHIP BUT RATHER A MONOTONIC RELATIONSHIP (IN THE SAME DIRECTION). IT DOES SO BY COMPARING THE RANK OF THE POINTS

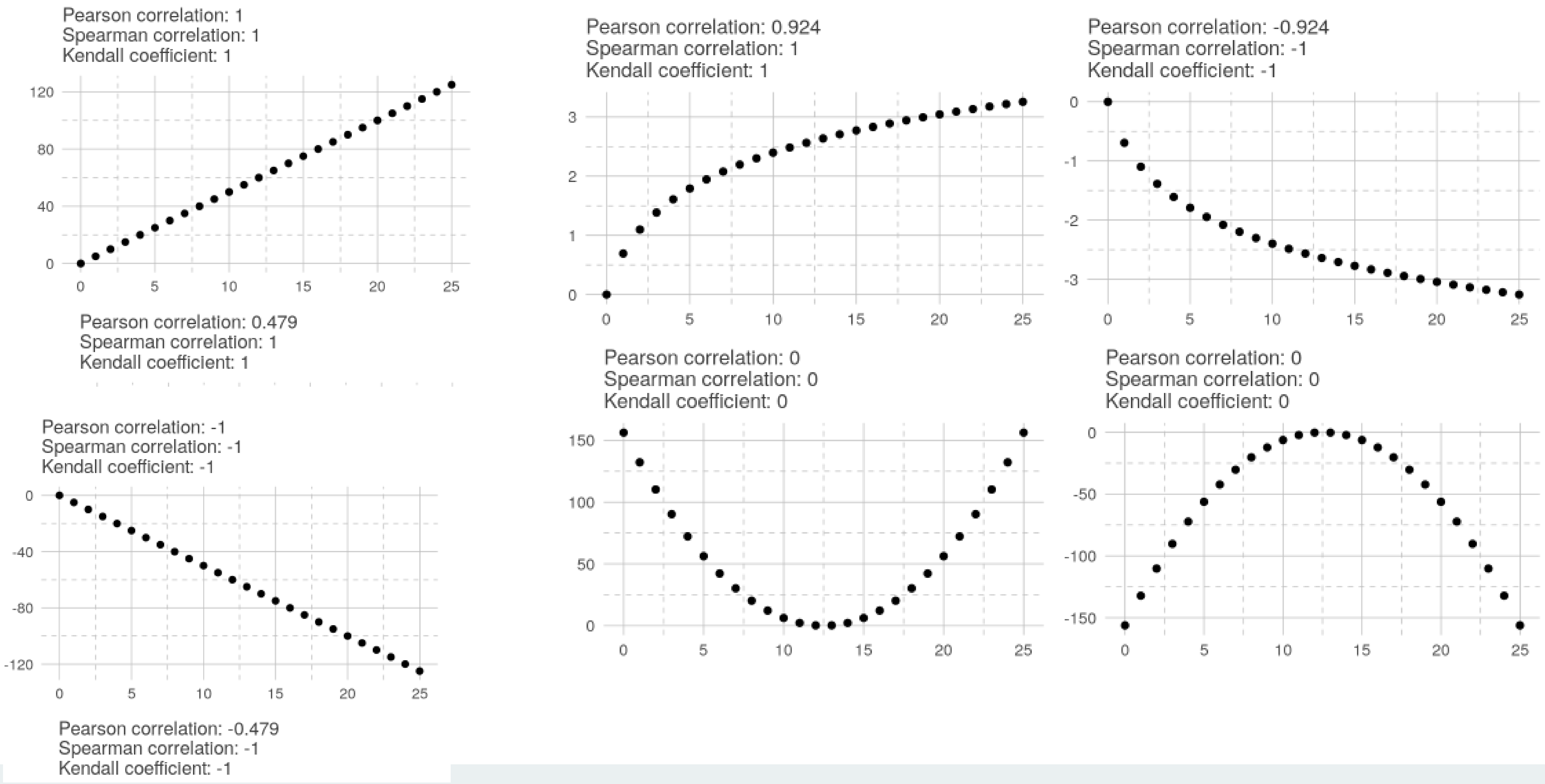


$$\tau = \frac{(\text{number of concordant pairs}) - (\text{number of discordant pairs})}{\binom{n}{2}}. \text{ [3]}$$

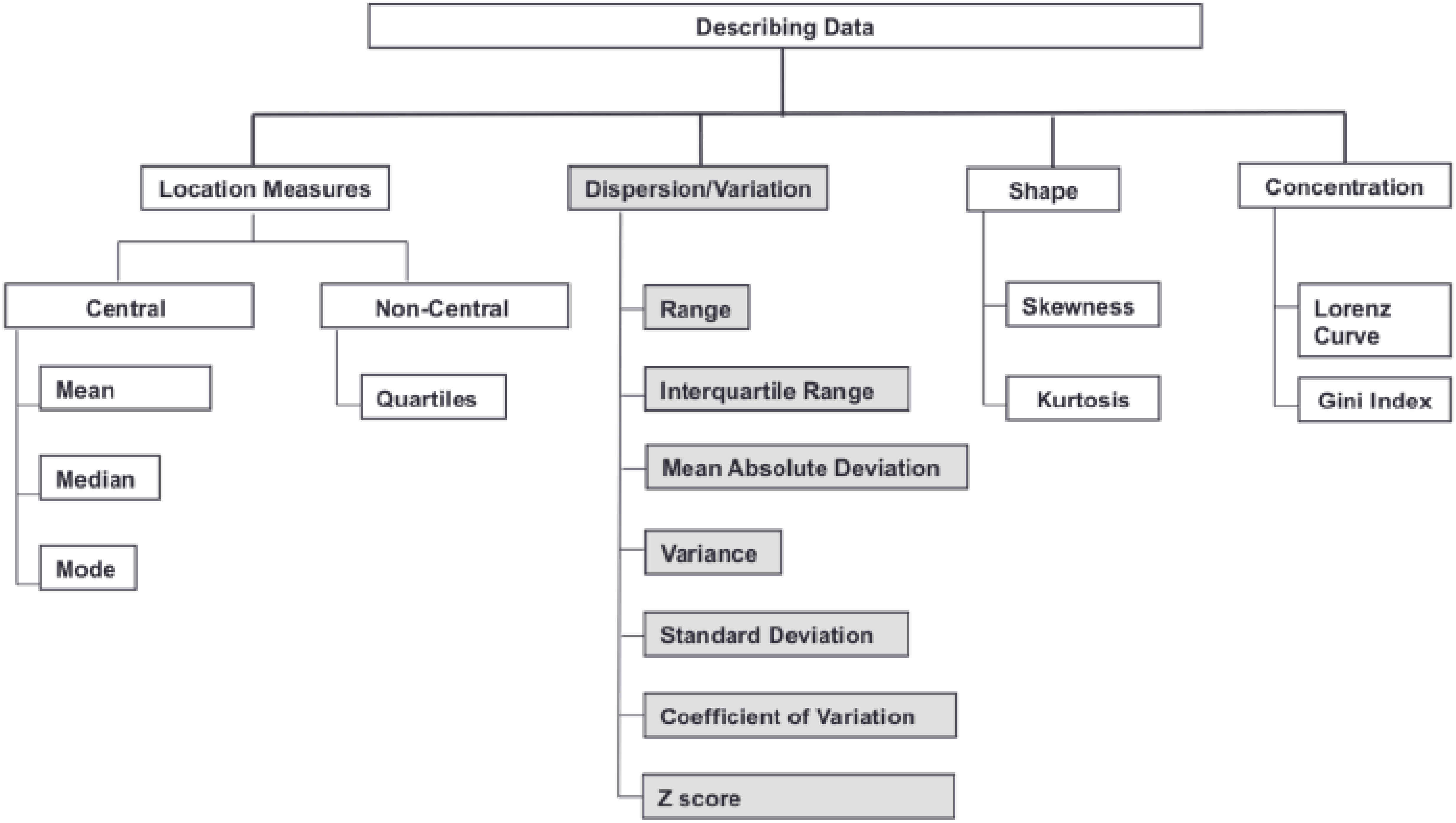
Where $\binom{n}{2} = \frac{n(n-1)}{2}$ is the binomial coefficient for the number of ways to choose two items from n items.

~KENDALL TAU PICKS UP ALL PAIRS OF OBSERVATIONS AND SEES IF THERE ARE MORE OBSERVATIONS "AGREEING" OR "DISAGREEING" ON THE DIRECTION OF GROWTH

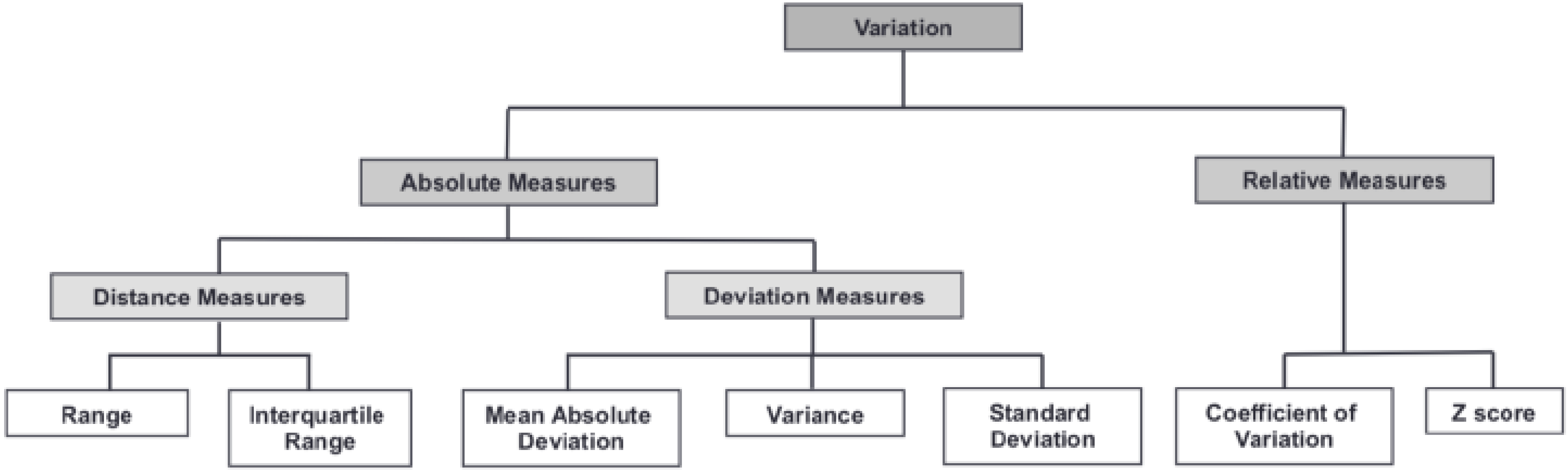
CORRELATION COMPARISONS



A SUMMARY


Describing Data
Location Measures
Dispersion/Variation
Shape
Concentration
Central
Non-Central
Mean
Median
Mode
Quartiles
Range
Interquartile Range
Mean Absolute Deviation
Variance
Standard Deviation
Coefficient of Variation
Z score
Skewness
Kurtosis
Lorenz Curve
Gini Index

A SUMMARY - VARIABILITY



REFERENCES

CHAPTER 2, NEWBOLD, CALSON, THORNE, STATISTICS FOR BUSINESS & ECONOMICS

https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

https://towardsdatascience.com/do-you-have-a-trustworthy-gut-the-most-counterintuitive-probability-problems-7b76aff941cb

ANY QUESTIONS ?