Quantitative Methods

Summary Data

Central Tendency : Center of data

Measures of Dispersion → Wide spread is data

- These measures indicate the extent of variation in the data set.
  - Standard deviation →
  - Range →
  - Interquartile range →
  - Percentiles →

Summary Data

Measures of Dispersion: Standard deviation

- Most commonly used measure of dispersion that measures deviations from the mean

distance

- Formula: Sample standard deviation:

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

population

- If sample data is used, then the divisor should be n -1 as shown below.

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

sample

Summary Data

Steps to calculate S.D. ✓ i) Calculate average

ii) For each value of x, reduce the mean → $x - \bar{x}$

iii) Square $(x_i - \bar{x})$

iv) Total $(x_i - \bar{x})^2$

v) Divide $\frac{(x_i - \bar{x})^2}{n}$

vi) Take square root of above $\left(\frac{39350}{5}\right)^{1/2}$

Measures of Dispersion: Standard deviation: Calculation

- Example:

CALCULATION OF STANDARD DEVIATION

(1) x_i	(2) $x_i - \bar{x}$	(3) $(x_i - \bar{x})^2$
75	−65	4,225
120	−20	400
280	140	19,600
195	55	3,025
30	−110	12,100
Mean = 140	Total = 0	Total = 39,350

-65^2

-20^2

$\bar{x}$

$$\text{Standard deviation} = \sqrt{\frac{39{,}350}{5}} = 88.71p$$

Summary Data

Calculate S.D from following data

8, 13, 16, 10, 3

Summary Data

Measures of Dispersion: Standard deviation

- Standard deviation when frequency distribution is given is calculated as

$$\text{Standard deviation} = \sqrt{\frac{\sum_{i=1}^{m} (x_i - \bar{x})^2 f_i}{n}}$$

Summary Data

S^2 → Variance is square of S.D.

σ^2 → Variance is square of S.D.

S.D. is square root of Variance

↳ S or σ

Measures of Dispersion: Standard deviation

- Example: Standard deviation when frequency distribution is given:

CALCULATING THE STANDARD DEVIATION FOR DISCRETE DATA IN TABULAR FORM

x_i	f_i	$(x_i-\bar{x})$	$(x_i-\bar{x})^2$	$(x_i-\bar{x})^2 f_i$
0	100	−1.27	1.6129	161.29
1	80	−0.27	0.0729	5.83
2	70	0.73	0.5329	37.30
3	40	1.73	2.9929	119.72
4	10	2.73	7.4529	74.53
$\bar{x}=1.27$	300			Total = 398.67

$\frac{\Sigma x_i f_i}{n}$

$$\text{Standard deviation} = \sqrt{\frac{398.67}{300}} = 1.15$$

Summary Data

SD: Three types of questions
i) No freq. distrib ii) x_i & freq. is given
iii) Intervals & freq.
↳ Calculate mid pts.

Measures of Dispersion: Standard deviation

- Example: Standard deviation for a grouped data

CALCULATING THE STANDARD DEVIATION FOR GROUPED DATA

Share price (pence)	Number of shares (f_i)	Midpoint (m_i) x_i	$m_i-\bar{x}$	$(m_i-\bar{x})^2 f_i$
Less than 20p	80	10	−46.6	173,725
20p or more, but less than 40p	120	30	−26.6	84,907
40p or more, but less than 60p	390	50	−6.6	16,988
60p or more, but less than 80p	210	70	13.4	37,708
80p or more, but less than 100p	200	90	33.4	223,112
Total	1,000			536,440

Mean = 56.6

$$\text{Standard deviation} = \sqrt{\frac{536,400}{1,000}} = \sqrt{536.44} = 23.16\text{p}$$

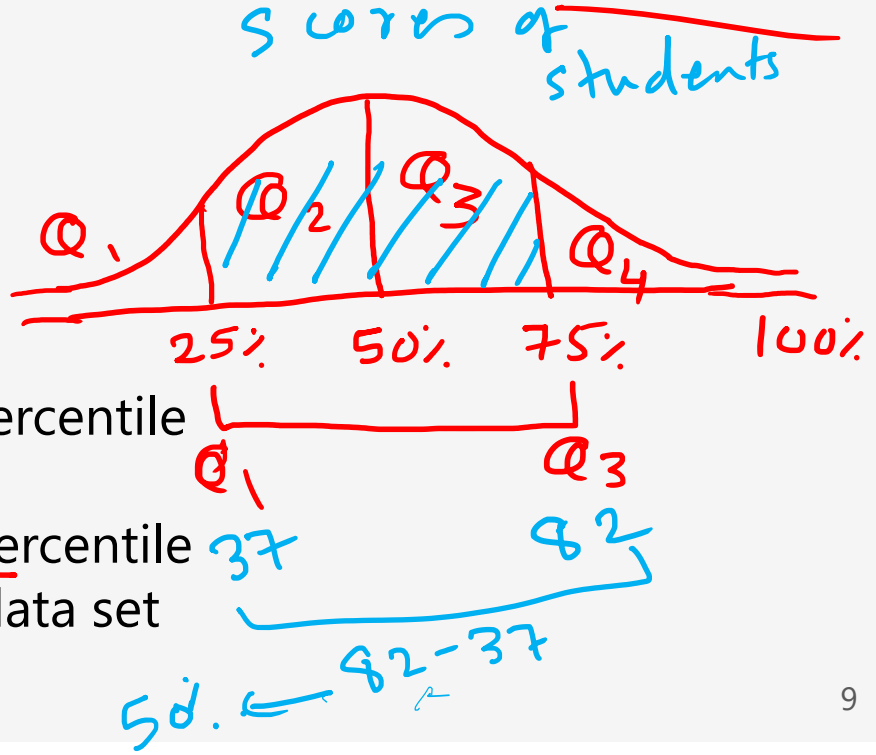
Summary Data

Measures of Dispersion: Range

- Easiest measure of dispersion, Range is Max Value less Min Value
- Range is totally dominated by extreme values, and we cannot know anything about other values

Measures of Dispersion: Interquartile Range

- Quartiles distribute the data in four equal parts.
- 1st quartile has 25% value below it, it is also called 25th percentile
- 2nd quartile has 50% value below it, it is the **Median**
- 3rd quartile has 75% value below it, it is also called 75th percentile
- Interquartile range is Q3 – Q1 i.e. middle 50% values of data set



Summary Data

Measures of Dispersion: Interquartile Range: Example

- Suppose the data is as follows:

Range = 120 − 8 = 112

8, 14, 28, 37, 42, 43, 57, 60, 62, 68, 71, 79, 84, 87, 90, 92, 94, 99, 101, 120

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20

Q2 (the median) is the average of the 10th and 11th observations, i.e. (68p + 71p) ÷ 2 = 69.5p. Q1 is the first quartile, the average of the 5th and 6th observations (42.5p). Q3 is the third quartile, the average of the 15th and 16th observations (91p).

50% of value are below & above 69.5

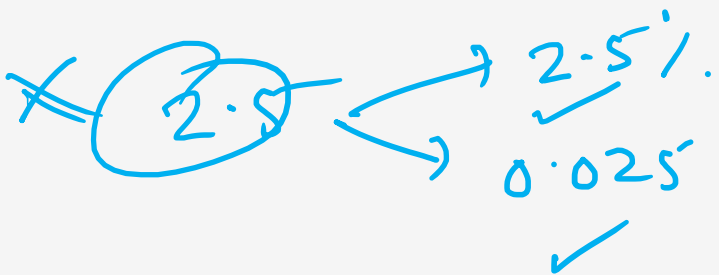
Q_1 : 42.5 Q_3 : 91

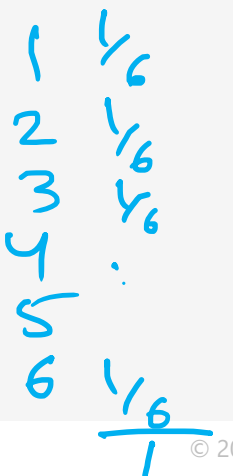
$Q_3 - Q_1$: 91 − 42.5 :

Summary Data

Probability Distributions

- Probability distribution shows the probabilities associated with all outcomes.
- For example, when we roll a fair die, there are six possible outcomes. All are equally likely.
- Hence probability of getting any number on a die is 1/6.
- Probability of any uncertain outcome lies between 0 and 1. 0 probability means it will never occur. Probability of 1 means it will surely occur.
- Sum of Probabilities of all outcomes always equals 1.



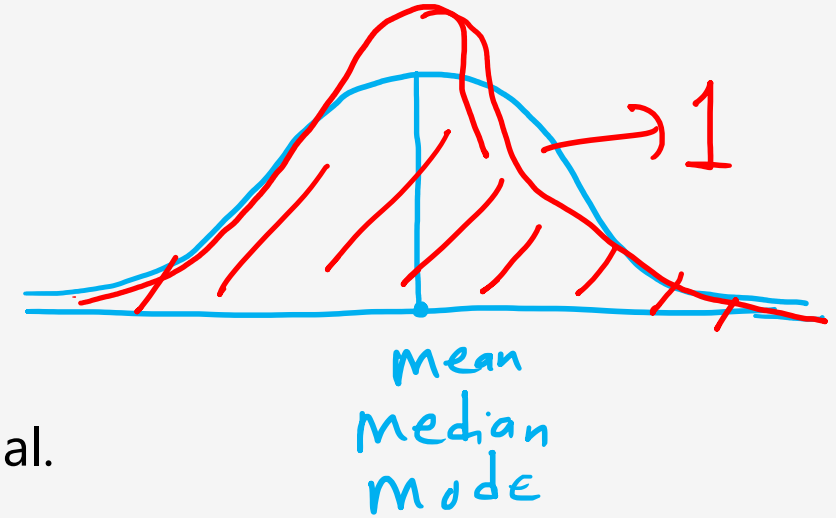


Summary Data

distrib A : Mean : 10%. SD : 20%. → Normal
B : Mean : 8%. SD : 12%. → Normal

Probability Distributions: Normal Distribution

- Normal distribution is most widely studied and used.



Properties of Normal Distribution

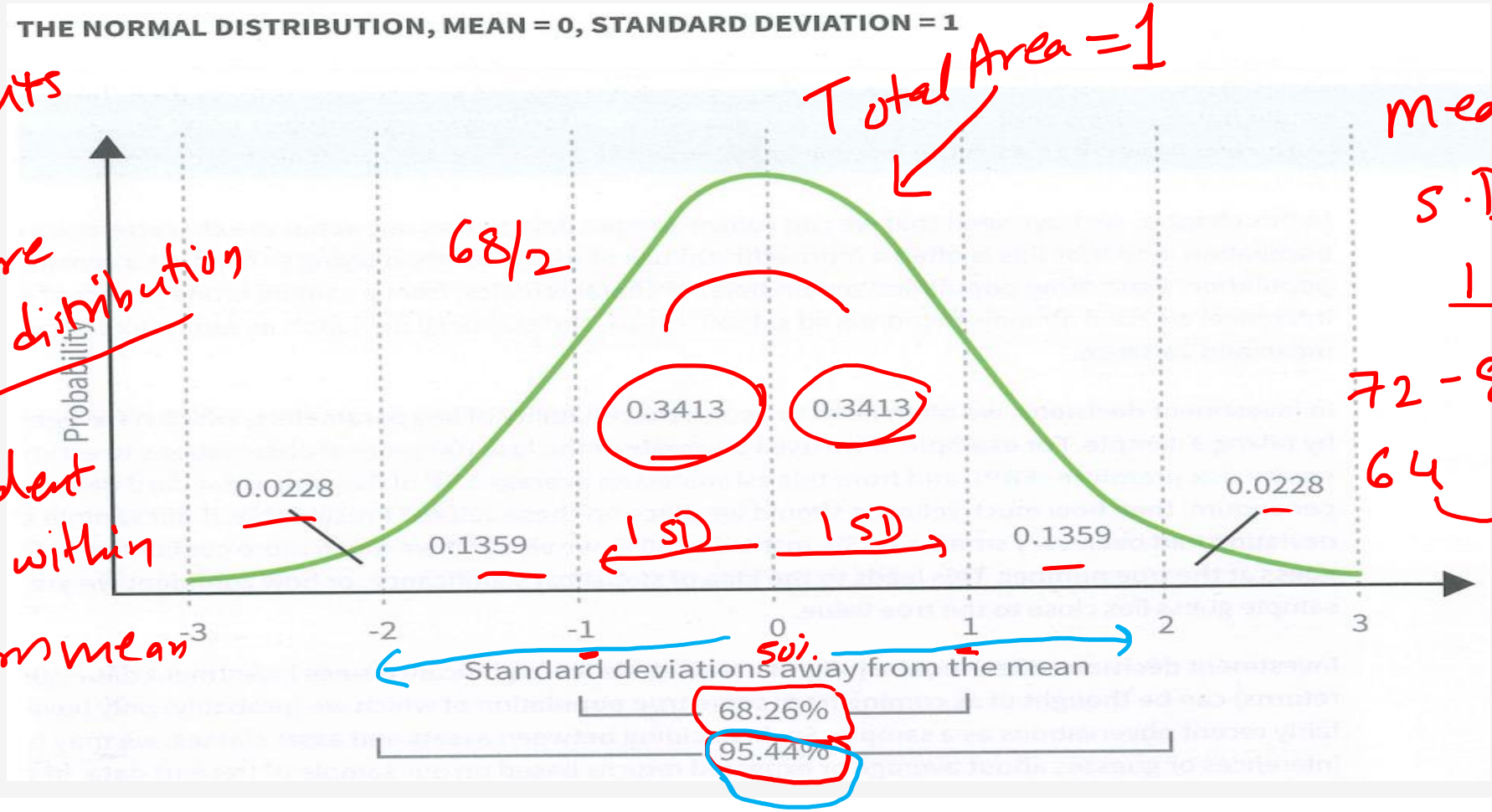
- Perfectly symmetrical, both side of the distribution are equal.
- Zero skewness. Mean = Median = Mode
- Shape of normal distribution depends only on two variables: Mean and Standard Deviation
- Sum of Probabilities of all outcomes always equals 1.

(1) (2)

Summary Data

1 SD away from mean : 68%.
2 SD " : 95%.
3 " total prob 99%.

Probability Distributions: Normal Distribution



100 students exam scores are normally distribution

68% student scored within 1 SD from mean

Total Area = 1

68/2

mean: 72

S.D: 8

1 SD range

72-8 64

72+8 80

68%

50%

Summary Data

Probability Distributions: Normal Distribution

Assume marks of students are normally distributed.

Mean: 72 S.D: 8.

i) Given above info, what proportion of students have scored marks betw 64 & 80.

72 − 8 = 64 72 + 8 = 80 : 68%.

ii) 68% of students must have scored between which range of marks?

1 SD

72 − 8 = 64 72 + 8 = 80

iii) What is the proportion of area bet Mean & 1 SD below mean? → 68%/2 → 34%.

Summary Data

Probability Distributions: Normal Distribution

Assume marks of students are normally distributed.

Mean: 72 S.D: 8.

i) Given above info, what proportion of students have scored marks betw 56 & 88.

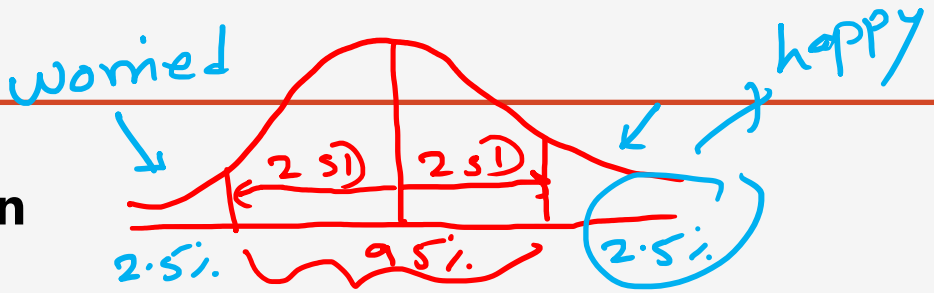
72 − 16 = 56 72 + 16 = 88 : 68%. 95%.

ii) 95% of students must have scored between which range of marks?

2 SD

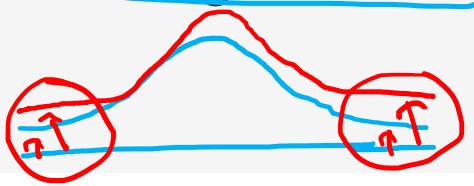
72 − 16 = 56 72 + 16 = 88

Summary Data



Probability Distributions: Normal Distribution

- If we assume returns on any asset such as stocks is normally distributed, then probability of extreme returns is quite low.
- For example, risk that return will be two standard deviations below the mean is only about 2.5% or in 5 out of 200 days.
- Financial crises of 2008 had days for which losses were 25 standard deviations below the mean.
- This shows that in real life returns are not normally distributed.
- In practice, the probability of extreme events is higher than what normal distribution tell us. This is known as problem of 'fat tails'.

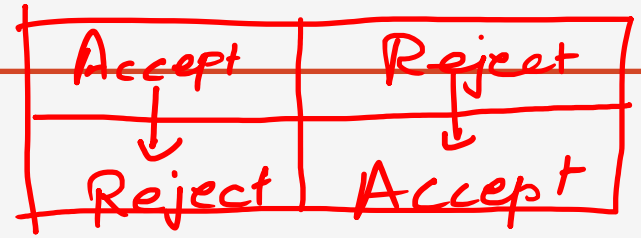


Summary Data

Statistical Significance and Investment Decisions

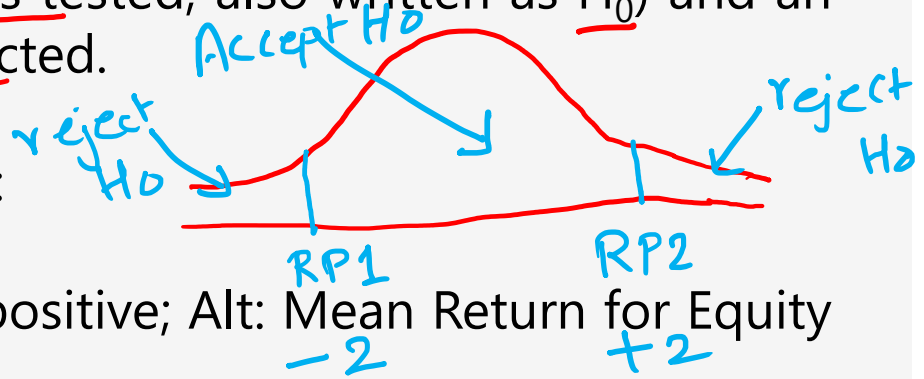
- If often use sample data to conclude about a population.
- In investment world, we use sample data which could be historical data to make future predictions.
- It is important to talk about how confident we are about the estimates. This is the idea behind statistical significance.
- Thus, when we make a statement about the population, we are making a hypothesis. This hypothesis should be tested at a given level of statistical significance.

Summary Data



Statistical Significance and Investment Decisions

- We formulate two hypotheses, a Null hypothesis (which is tested, also written as H_0) and an alternative hypothesis (H_1) which is accepted if Null is rejected.



Example of Null and Alternative hypothesis can be as follows:

- Null: Mean Return for Equity funds is equal to zero or is positive; Alt: Mean Return for Equity funds is negative

- To test the hypothesis, we calculate Test Statistic and compare it with critical value or rejection points.

- If Test statistic is more extreme than rejection points, we reject the Null and accept Alt hypothesis.
- Else, we fail to reject Null and reject Alt hypothesis.

Correlation and Bivariate Linear Regression

Correlation and Bivariate Linear Regression

Learning outcomes……..

- **Define** correlation and **Identify** alternative measures of correlation
- **Explain** the least-squares regression technique in deriving line of best fit and **Interpret** the correlation coefficient R, R squared, and adjusted R squared and measures of unexplained variation such as mean squared error
- **Calculate and Interpret** a forecast value of the dependent variable given the intercept and slope coefficients of a regression equation considering their statistical significance and adjusted R squared

Correlation and Bivariate Linear Regression

Learning outcomes…….. continued

- **Explain** the shortfalls in application of linear regression to forecasting, including why correlation does not imply causation and the pitfalls of data-mining

- **Describe** the impact of extreme events on alternative measures of correlation

Correlation and Bivariate Linear Regression

Measures of Correlation

- Rank Correlation coefficient
  - Rank correlation measures the strength of association between two ranked variables.
  - Denotes as r_s or ρ.
  - It tests whether higher values of one variable tend to be associated with higher values of another variable.

Correlation and Bivariate Linear Regression

i) Two parameters

ii) Rank companies on two parameters

iii) Diff & apply formula

Measures of Correlation: Rank Correlation coefficient

RANKED EARNINGS AND DIVIDEND GROWTH RATES OF A SAMPLE OF FIVE COMPANIES

Company	Earnings growth rate	Rank	Dividend growth rate	Rank	Difference (d_i)
A	10%	3	4%	4	-1
B	6%	4	9%	1	3
C	-3%	5	0%	5	0
D	14%	2	5%	3	-1
E	20%	1	6%	2	-1

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

$$\rho = 1 - \frac{6(12)}{5(5^2-1)}$$

$$\rho = 0.4$$

Correlation and Bivariate Linear Regression

Measures of Correlation

- Rank Correlation coefficient
  - Can take values between -1 and 1.
  - Value of +1 indicates perfect association (very strong positive relationship), value of 0 means no association and value of -1 means perfect negative (very strong negative) correlation.
- We can create hypothesis test for checking whether there is an association between two variables or not

Correlation and Bivariate Linear Regression

Measures of Correlation

- Simple Correlation coefficient:

$$\rho = \frac{\sum_i (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_i (X_i - \bar{X})^2 \sum_i (Y_i - \bar{Y})^2}} = r$$

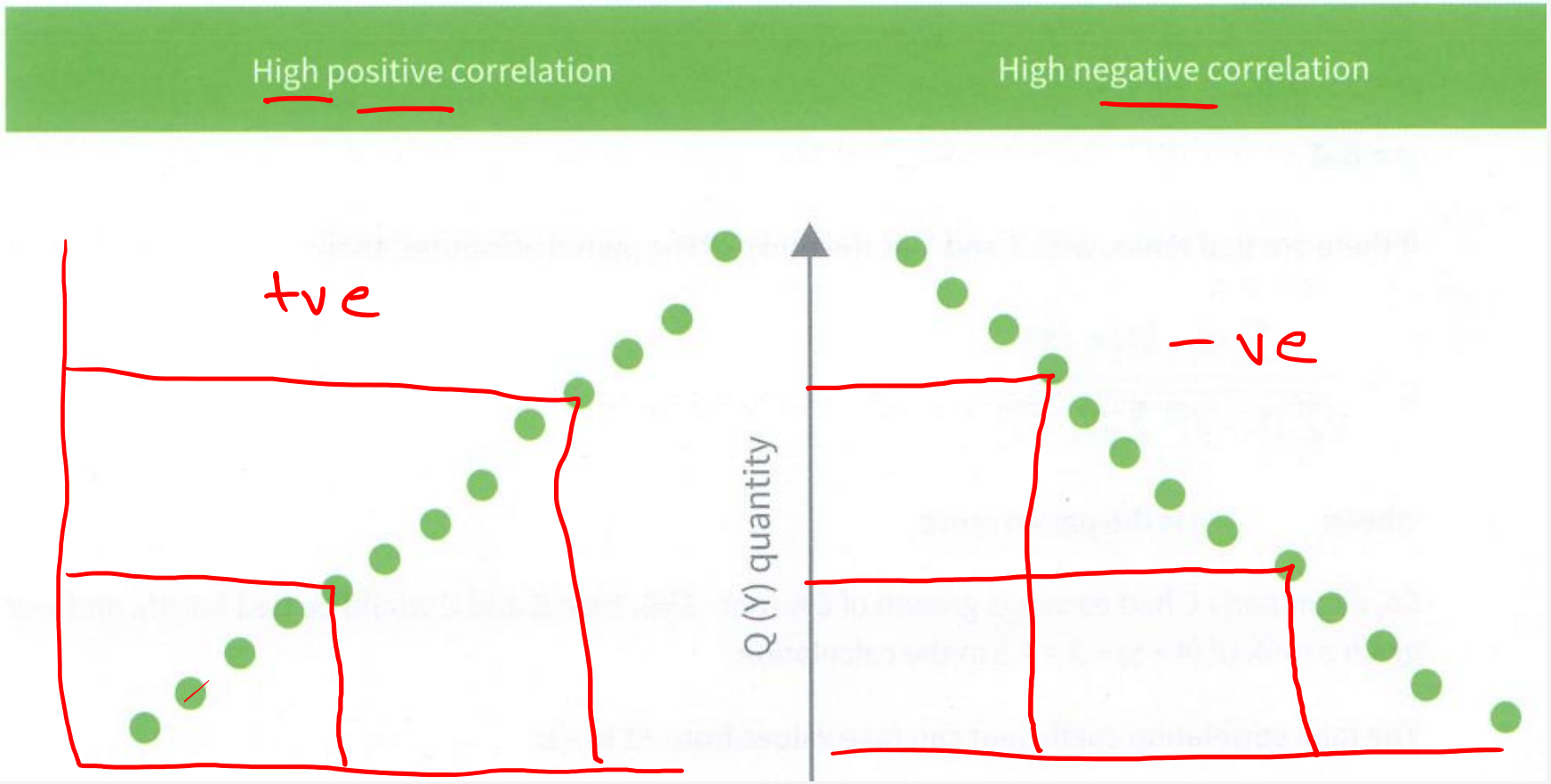
↳ widely used

- Value of +1 indicates perfect association (very strong positive relationship), value of 0 means no association and value of -1 means perfect negative (very strong negative) correlation.

Correlation and Bivariate Linear Regression

Measures of Correlation

Examples of data scatters with different degrees of correlation are presented below:



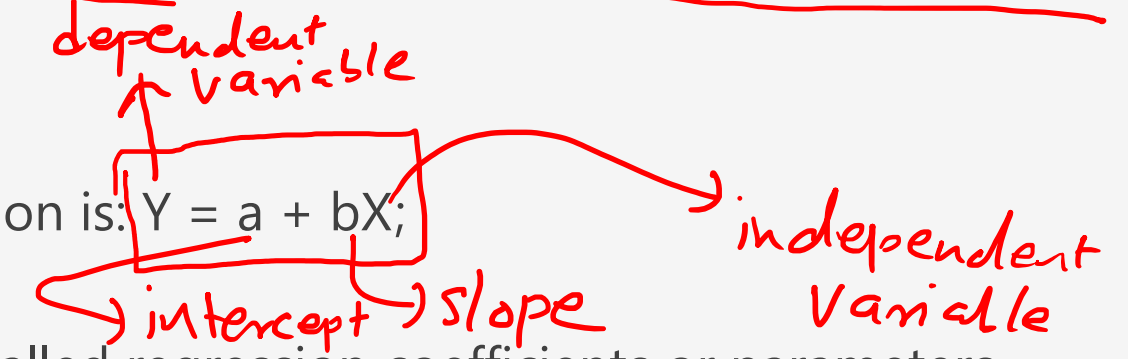
Correlation and Bivariate Linear Regression

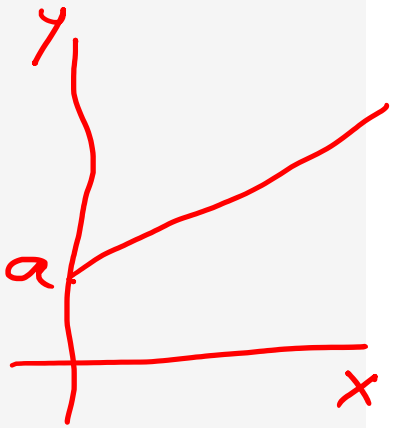
Measures of Correlation

Correlation and Bivariate Linear Regression

Regression Analysis: The least squared method

- Regression analysis helps us to estimate value of one variable (called dependent variable) from value of another one or many variables (called independent variable).
- If there is only one independent variable, it is called a Simple linear regression or Bivariate regression.
- The form of simple linear regression is: $Y = a + bX$;
  - a (intercept) and b (slope) are called regression coefficients or parameters
  - both the coefficients can be estimated from historical data for X and Y





Correlation and Bivariate Linear Regression

Regression Analysis: The least squared method: Scatter diagram

Correlation and Bivariate Linear Regression

Regression Analysis: The least squared method: Scatter diagram

- In the diagram, a straight line is drawn which indicates the relationship between Y and X in the best possible manner.
- It is unlikely to have all points on the line, hence the line should be fitted in a way that minimizes distance of points from the line.
- This is done with the help of least squares method, which minimizes sum of the errors squared.
- Errors are the distance of each point from the line.