Assessment and Analysis of Test Data

Presented by SARF

Presenter:
Ron Berkers

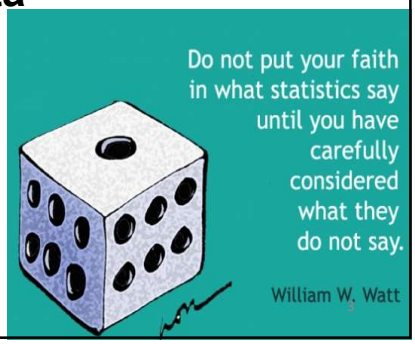
1

Module B
Statistical Methods

2

Statistics:

- Collection of data
- Organising of data
- Summarising of data
- Analysing of data
- Interpretation of data
- Presenting of data



Do not examine the whole group - called the population

Examine only part of the group – called the sample

Example:

You will not test the whole soil layer, you will take a sample consisting of 2 bags of 50 kg each. (Representative sample)

4

Central tendency

Average

- a measure of the "middle" or "typical" value of a data set
- thus a measure of central tendency

- Mean (or arithmetic mean)
- Median
- Mode

5

Central tendency

Mean:

- The mean is the average of the numbers.
- add up all the numbers, then divide by how many numbers there are.

$$\bar{X} = \frac{\sum_{j=1}^{n} x_j}{n}$$

Example: What is the Mean of these numbers?

6, 11, 7

- Add the numbers: 6 + 11 + 7 = 24
- Divide by *how many* numbers (there are 3 numbers): 24 / 3 = 8

The Mean is 8

- 6, 11 and 7 added together is the same as 3 lots of 8

6

Central tendency

Median:

- Defined as the "middle" value of the set when the values are arranged in order.

Example: Consider the following numbers:

{3, 13, 7, 5, 21, 23, 39, 23, 40, 23, 14, 12, 56, 23, 29}

- If we put those numbers in order we have:

{3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 39, 40, 56}

- There are fifteen numbers. Our middle number will be the eighth number:

{3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 39, 40, 56}

- The median value of this set of numbers is 23.

(Note that it didn't matter if we had some numbers the same in the list)

7

Central tendency

Example: Two Numbers in the Middle

What if there are an even amount of numbers?

In that case we need to find the middle pair of numbers, and then find the value that would be half way between them.
This is easily done by adding the middle pair together and dividing by two.

- Consider the following numbers:

{3, 13, 7, 5, 21, 23, 23, 40, 23, 14, 12, 56, 23, 29}

- If we put those numbers in order we have:

{3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 40, 56}

8

Central tendency

- There are now fourteen numbers and so we don't have just one middle number, we have a pair of middle numbers:

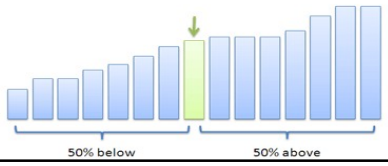
{3, 5, 7, 12, 13, 14, 21, 23, 23, 23, 23, 29, 40, 56}

- In this example the middle numbers are 21 and 23.
- To find the value half-way between them, add them together and divide by 2:

21 + 23 = 44
44 ÷ 2 = 22

- And, so, the Median in this example is 22.

Median



9

Central tendency

Mode:

- The Mode is the value that occurs most often.

Example:

- Consider the following numbers:

{13,13,2,13,2,2,13,2,13,2,13}

- Group the numbers so we can count them:

{2, 2, 2, 2, 2, 13, 13, 13, 13, 13, 13}

- "13" occurs 6 times, “2" occurs only 5 times
- The mode is 13.

10

Central tendency

Example:

- **Consider the following numbers:**

{3, 6, 4, 7, 6, 4, 5}

- **Group the numbers so we can count them:**

{3, 4, 4, 5, 6, 6, 7}

- **"4" and "6" both occur 2 times**
- **Both 4 and 6 are modes.**
- **This is called "Bimodal"**
- **3 or more modes is called "Multimodal"**

11

Central tendency

Quantile:

- **We often want to summarize a set of numbers in a few numbers, for ease of reporting or comparison. The most direct method is to use quantiles. The quantiles are values which divide the set of numbers such that there is a given proportion of observations below the quantile.**
- **For example, the median is a quantile. The median is the central value of a set of numbers, such that half the points are less than or equal to it and half are greater than or equal to it.**

12

Central tendency

Quartile:

- **Quartiles divide a set of numbers (distribution) into four equal parts.**

Example:

- **Consider the following numbers:**

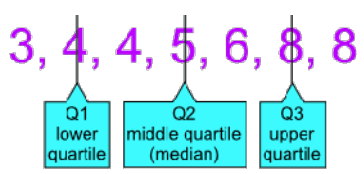
{5, 8, 4, 4, 6, 3, 8}

- **Put them in order:**

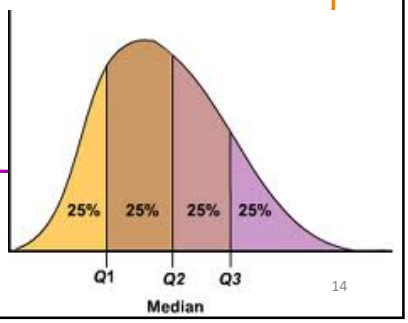
{3, 4, 4, 5, 6, 8, 8}

13

- **Cut the list into quarters:**



- **Q1 = 4**
- **Q2 = 5 (which is also the median)**
- **Q3 = 8**



14

Measures of Dispersion

Dispersion:

- Measures of central tendency are used to estimate "normal" values of a dataset
- Measures of dispersion are important for describing the spread of the data, or its variation around a central value
- Two distinct samples may have the same mean or median, but completely different levels of variability, or vice versa

Example:

10 homes with a value of R500000 have the same mean as 2 homes with a value of R750000 and 8 homes with a value of R125000.
The mean will not show the variation between the prices of the houses.

15

Measures of Dispersion

Range:

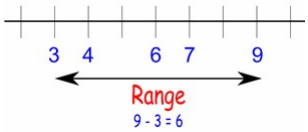
- The simplest measure of dispersion.
- The range is calculated by taking the difference between the maximum and minimum values in the data set.
- The range only provides information about the maximum and minimum values and does not say anything about the values in between.

Example:

- Consider the following numbers:

{4, 6, 9, 3, 7}

- The lowest value is 3, and the highest is 9.
- The range is: 9 – 3 = 6

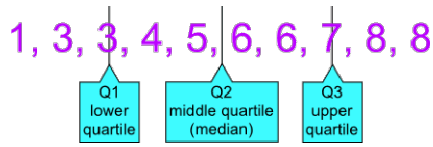


16

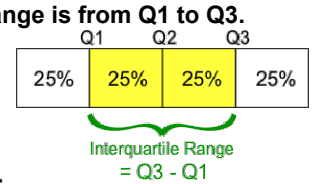
Measures of Dispersion

Inter – Quartile Range:

- A good indicator of the spread in the centre region of the data
- More resistant to extreme values than the range



- The inter-quartile range is from Q1 to Q3.



- Quartile 1 (Q1) = 3
  Quartile 2 (Q2) = 5.5
  Quartile 3 (Q3) = 7

- The Inter-quartile Range is:

$$Q3 - Q1 = 7 - 3 = 4$$

17

Measures of Dispersion

Mean Deviation:

- Uses each data value
- It is the mean of the distances between each value and the mean
- It gives us an idea of how spread out from the centre the set of values is

$$\overline{Dev} = \frac{\sum_{j=1}^{N} |x - \bar{x}|}{N}$$

Example:

- Consider the following numbers:

$$\{9,2,7,5,4,6,5,2\}$$

- The mean of these values is:

$$(9+2+7+5+4+6+5+2)/8 = 5$$

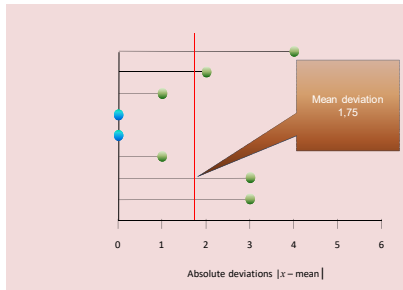
- Each value can now be expressed as a deviation from the mean by subtracting the mean from it:

$$(9-5,\ 2-5,\ 7-5,\ 5-5,\ 4-5,\ 6-5,\ 5-5,\ 2-5)$$
$$= (4,\ -3,\ 2,\ 0,\ -1,\ 1,\ 0,\ -3)$$

18

- Negative signs are ignored (i.e. the absolute deviations are considered), hence:

{4 3 2 0 1 1 0 3}



- The set of values are the original data expressed as absolute deviations from the mean. The mean of these absolute deviations can be calculated as:

(4+3+2+0+1+1+0+3)/8 = 1,75

19

Measures of Dispersion

Variance:

- The average of the squared differences from the Mean.
- Not often used.

$$\sigma^2 = \frac{\sum_{j=1}^{N}(x - \bar{x})^2}{N}$$

Example:

- Consider again the following numbers:

{9,2,7,5,4,6,5,2}

- The variance for this set is:

$$\sigma^2 = (4^2+(-3)^2+2^2+0^2+(-1)^2+1^2+0^2+(-3)^2)/8 = 5$$

20

Measures of Dispersion

Standard Deviation:

- Standard Deviation is defined as the square root of the variance.

$$\sigma = \sqrt{\frac{\sum_{j=1}^{N}(x - \bar{x})^2}{N}}$$

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

Use σ for population and s for a sample

21

Measures of Dispersion

Example:

- Consider again the following numbers:
  {9,2,7,5,4,6,5,2}
- The standard deviation for this set is:

$$\sigma = \sqrt{(4^2+(-3)^2+2^2+0^2+(-1)^2+1^2+0^2+(-3)^2)/8} = 2{,}236$$

- Consider a sample of this set:
  {9,2,7)
- The standard deviation of the sample of this set is:

$$s = \sqrt{(4^2+(-3)^2+2^2/(3-1)} = 3{,}81$$

22

Measures of Dispersion

Example:

Sampling of concrete

- Concrete must be sampled from that being used for construction (according to SANS 5861-2)
- Each sample (sufficient for the slump test and three test cubes) must be taken from a different batch of concrete chosen on a random basis.
- At least one sample must be taken from each day's placing and from at least every 50 m³ of concrete of each grade being placed.
- When the batch size is less than 2 m³ , the first sample of each grade shall be taken after at least three batches of this grade have been mixed and discharged.

23

Measures of Dispersion

Acceptance Criteria:

For a small numbers of result (number of values less than 30):

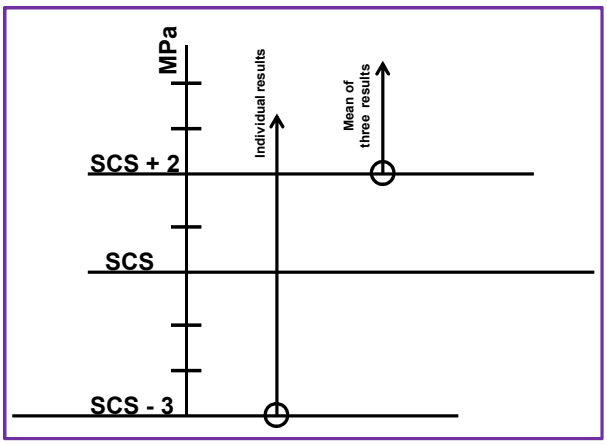
Strength test results shall meet <u>both</u> of the following acceptance criteria:

- No individual result (average of three cube strength) shall be more than 3 MPa below the specified characteristic strength (SCS).
- The mean of any group of three consecutive and overlapping results shall exceed the specified characteristic strength by at least 2 MPa.

24

Measures of Dispersion

Acceptance criteria for an individual result and mean of three consecutive results:



25

Measures of Dispersion

Acceptance Criteria:

For a large numbers of result (number of values more than 30):

- **The average of overlapping sets of 30 results for a specific grade of concrete shall exceed specified strength by 1,7 times the standard deviation.**
- **No individual result shall be more than 3 MPa below the specified characteristic strength.**

26

Measures of Dispersion

Process Control:

- **Specified Characteristic Strength (SCS) is defined as that strength below which no more than 5% of the valid results may fall.**
- **To ensure that no more than 5% of results are below SCS, a higher strength should be aimed at. This strength is called Target Average Strength (TAS).**
- **Roughly half of the valid results should be above TAS, and half below.**
- **To meet the above criteria, TAS should be 1,645 standard deviations above SCS.**

27

Measures of Dispersion

Process Control:

- **The standard deviation of the results will depend on the degree of control achieved during production. The better the control, the lower the standard deviation and, therefore, a lower TAS may be targeted.**
- **Below are given standard deviations and margins for different degrees of control as recommended by some experts as starting points for use at the beginning of a project:**

Degree of control	Standard deviation, Mpa	Margin, Mpa
Poor	7	11,5
Average	6	10,0
Good	5	8,5

28

Measures of Dispersion

Examples of strength result data: Specified Characteristic Strength is 35 MPa:

Test Number	Result, Mpa	Test Number	Result, Mpa
1	58	16	45
2	56	17	44
3	57	18	44
4	45	19	45
5	48	20	52
6	46	21	46
7	47	22	46
8	45	23	48
9	40	24	53
10	46	25	51
11	50	26	49
12	50	27	48
13	44	28	39
14	48	29	45
15	39	30	50
		Sum:	1424

29

Measures of Dispersion

For the previous results:

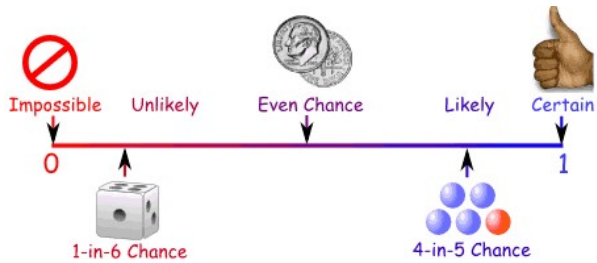
- **Mean = $\bar{x}$ = 1424/30 = 47,5 MPa**
- **Standard Deviation = s = 4,7 MPa (all results – mean, sum of squared numbers/(30-1), square root of previous result)**
- **The value is smaller than the value of 6 taken for average control.**
- **The results therefore show a degree of control closer to "good" than to "average".**
- **The margin, to the nearest 0,5 MPa, should therefore be 1,7 x 4,7 = 8 MPa and not 10 MPa.**
- **TAS should be 43,0 MPa (minimum $\bar{x}$ = SCS + 8 = 35 + 8 = 43 MPa).**
- **According to this assessment, average strength is 47,5 – 43 = 4,5 MPa higher than required.**
- **Concrete to strong, reduce cement content (increase w:c), more economical**

30

Probability

Probability P(*A*):

- **Probability is the chance that something will happen - how likely it is that some event will happen.**
- **A probability is a way of assigning every event a value between zero and one, with the requirement that the event made up of all possible results is assigned a value of one**



31

Probability

Independent probability:

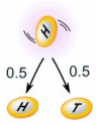
Calculated when the events are not affected by the previous events.

P(*A* and *B*) = P(*A*) x P(*B*)

Example:

What are the chances of the side of two coins being both heads?

Probability = ½ x ½ = ¼



32

Probability

Mutually exclusive events:

Mutually Exclusive means you can't get both events at the same time.

$$P(A \text{ or } B) = P(A) + P(B)$$

Example:

What are the chances of rolling a 1 or 2 on a six-sided dice?

$$\text{Probability} = P(1 \text{ or } 2) = P(1) + P(2)$$

$$= \frac{1}{6} + \frac{1}{6}$$

$$= \frac{1}{3}$$

33

Probability

Not mutually exclusive events:

Not Mutually Exclusive means you can get both events at the same time.

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Example:

When drawing a single card at random from a regular deck of cards, what are the chances of getting a heart or a face card (J,Q,K), or one that is both?

$$\text{Probability} = P(H \text{ or } F) - P(H \text{ and } F)$$

$$= \frac{13}{52} + \frac{12}{52} - \frac{3}{52}$$

$$= \frac{11}{26}$$

52 cards; 13 are hearts; 12 are face cards; 3 are hearts and face cards

34

Probability

Conditional Probability:

Conditional probability means that one event only happens after another has happened.

Probability = P(A) x P(*B*|*A*)

Example:

2 blue and 3 red marbles are in a bag. What are the chances of getting a blue marble first (event A(and second (event B)?

Probability of getting blue marble first (P(*A*): $\frac{2}{5}$

Blue marble first, chances of getting another blue marble (P(B|A): $\frac{1}{4}$

If we got a red marble first, chances of getting a blue marble: $\frac{2}{4}$

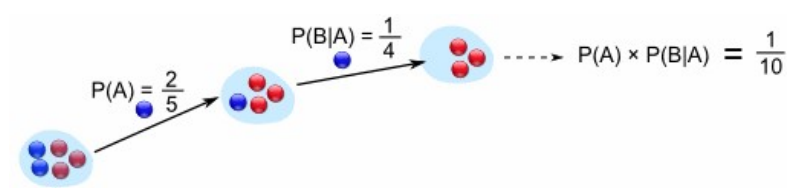
35

Probability

Conditional Probability = P(A) x P(*B*|*A*)

$$= \frac{2}{5} \times \frac{1}{4}$$

$$= \frac{1}{10}$$



36

Sampling

Random Sampling (Probability Sampling):

- **Test points must be chosen by unpredictable means.**
- **Samples must be chosen so as to be representative of a population.**
- **A fundamental characteristic of random sampling is that each member of a population has an equal chance of being included in the sample.**
- **Acceptance sampling is used to determine if a production lot of material meets the governing specifications.**

37

Sampling

Simple random Sampling:

- **The simplest method of probability sampling.**
- **Within a particular study population everyone has an equal chance of inclusion in the sample.**
- **Considered 'fair' and therefore allows findings to be generalized to the whole population from which the sample was taken.**
- **Applicable when population is small, homogeneous & readily available.**
- **A lottery system or table of random numbers is used to determine which units are to be selected.**

	1 2 3 4	5 6 7 8	9 10 11 12	13 14 15 16	17 18 19 20	21 22 23 24	25 26 27 28	29 30 31 32
1	8 0 9 4	2 5 2 6	8 2 4 7	1 3 4 7	7 4 3 3	3 6 2 0	1 8 9 7	2 1 3 4
2	3 5 6 3	2 1 9 8	8 2 1 1	9 0 4 5	2 6 1 6	2 7 5 1	2 5 2 7	1 0 9 5
3	1 3 3 0	6 3 1 1	3 7 5 3	9 0 9 3	8 7 3 8	6 8 1 5	1 5 3 8	5 5 4 3
4	3 5 6 5	0 0 1 6	5 2 4 3	6 4 9 2	4 7 9 8	6 0 9 5	5 2 8 3	1 6 2 0
5	7 5 5 0	5 9 2 5	5 5 8 8	7 3 1 1	2 1 9 2	4 5 4 5	3 5 3 0	5 5 9 0
6	4 4 9 0	5 4 1 7	9 7 2 7	6 1 5 3	5 9 8 1	4 8 7 8	9 9 8 0	9 8 7 7
7	6 5 4 5	9 1 0 4	9 3 1 8	6 8 1 9	7 5 3 7	2 7 8 5	9 3 7 9	2 4 4 5
8	9 5 2 6	5 9 9 5	1 2 1 6	9 7 5 3	9 2 2 3	5 6 5 8	2 9 4 4	2 8 9 9
9	4 6 8 6	4 8 2 0	7 5 5 4	0 6 1 2	9 6 8 9	4 2 5 1	9 1 3 8	1 7 0 9
10	6 4 9 8	7 5 1 9	0 4 7 4	7 8 1 8	6 5 3 2	9 5 8 3	9 0 7 2	4 0 9 0
11	6 7 2 2	9 8 6 6	9 3 6 1	7 5 7 5	4 8 9 3	1 3 1 5	9 6 7 9	8 5 3 4
12	9 7 4 8	5 9 3 2	5 1 1 5	2 7 2 1	0 0 3 3	9 3 0 3	9 7 1 3	4 0 1 2
13	5 6 4 1	1 4 5 7	1 4 1 9	7 4 9 4	8 1 9 5	7 3 0 8	1 2 1 8	5 0 3 9
14	7 4 4 4	9 2 0 0	8 8 4 0	5 8 6 2	4 3 9 6	3 9 4	9 1 9 9	9 3 3 6
15	8 2 7 9	3 0 1 9	4 6 7 2	3 7 4 3	3 9 7 9	4 6 8 9	9 0 2 1	6 9 9 0
16	0 1 6 1	7 6 1 7	1 0 2 4	2 3 5 7	2 8 9 1	5 6 7 7	1 5 8 5	2 4 5 2
17	7 3 5 8	9 7 5 9	7 5 5 6	6 6 2 4	9 9 7 7	2 0 0 8	5 5 9 6	9 7 4 0
18	7 9 3 0	4 7 1 4	8 8 9 5	2 9 1 9	1 8 0 4	4 0 4 4	1 0 3 4	2 5 9 7
19	9 8 8 7	4 3 1 6	8 5 2 8	4 5 3 5	8 4 3 0	5 2 7 0	9 0 0 5	0 7 8 8
20	1 2 8 1	2 5 1 6	8 5 6 9	2 3 1 0	3 9 3 9	8 7 0 3	9 8 4 1	0 3 5 3
21	3 9 4 7	4 9 3 7	7 6 3 4	2 5 4 3	6 2 3 9	7 4 5 5	2 0 5 5	7 7 9 5
22	4 5 5 0	8 1 0 3	1 2 5 0	2 3 0 4	1 1 3 8	9 7 8 8	9 1 4 4	4 5 2 6
23	1 3 4 4	9 6 9 7	2 3 8 3	6 9 7 6	6 2 5 1	4 2 0 1	2 0 3 8	6 5 5 2
24	8 9 7 6	5 8 2 3	8 4 8 7	0 4 5 0	3 1 0 5	9 1 6 6	2 7 1 7	7 6 0 1
25	7 7 1 0	9 9 4 3	6 9 7 8	8 2 7 3	9 7 1 4	9 7 0 0	1 5 6 6	2 8 4 9
26	8 9 5 9	6 0 0 8	8 4 4 2	2 2 8 2	1 5 2 4	2 6 1 7	5 8 1 8	0 0 8 1
27	7 9 4 1	2 3 1 2	2 4 3 1	6 7 0 2	9 9 8 4	3 4 6 9	3 0 8 5	4 7 6 2
28	2 2 6 4	0 8 9 6	9 1 0 7	8 5 4 2	7 3 1 9	3 7 8 2	1 0 6 8	9 5 7 4
29	9 5 9 4	7 4 1 8	8 9 3 5	6 0 4 5	1 1 8 3	5 9 1 6	9 5 9 9	1 1 4 3
30	4 6 1 5	5 2 4 5	6 3 5 5	3 2 0 5	5 1 0 3	3 6 5 0	1 1 6 5	5 1 3 3

38

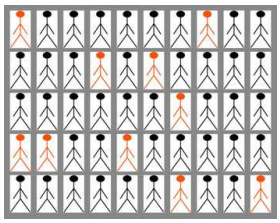
Sampling

Lottery method

- Population is finite i.e. number is relatively small and known
- Elements of population can be easily identified and numbered.

Example:

- 50 people in a class, need to draw a sample of 10 students.
- 50 numbers on paper in a hat, draw 10 numbers.



An example of simple random sampling of 10 subjects, represented by the red 'stickmen', selected at random from a total of 50 subjects using a lottery method.

39

Sampling

Table method

- Population is large
- Use tables of random numbers.
- Tables 7(a) to 7(c) in notes.

Example:

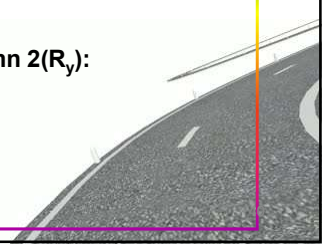
- Section of road to be tested: km 21,654 to 21,704, width is 3 meters.

- Start point: P_s = 21,654 km
  End point: P_e = 21,704 km
  Section length: $L_s = P_e - P_s$ = 21,704 – 21,654 = 0,050 km.
  Width: W_s = 3 m

- From table 7(a): Row 20, Column 1(R_x) and column 2(R_y):

R_x = 0,0941

R_y = 0,8008



Sampling

Example (Continued):

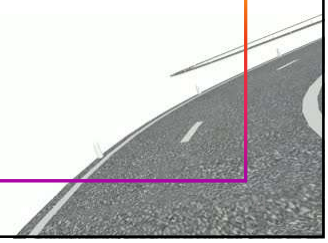
- R_x represents random number for the width coordinate.
  R_y represents random number for the length coordinate.

- L = km length were sample must be taken
  $= P_s + R_y L_s$
  = 21,654 +0,8008(0,050)
  = 21,694

- W = width = width from edge of road where sample must be taken
  $= R_x W_s$
  = 0,0941(3)
  = 0,282 m

- Next sample point:

  Same calculations, use row 21 in table 7(a)



Characteristics of Variability

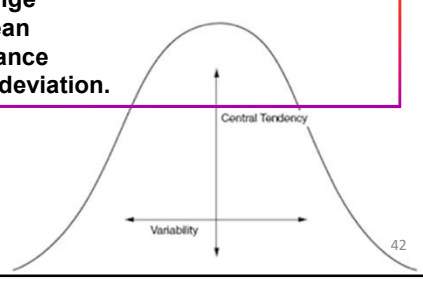
Variability:

- The extent to which data points in a statistical distribution or data set diverge from the average or mean value.

- Variability also refers to the extent to which these data points differ from each other.

- There are four commonly used measures of variability:

  Range
  Mean
  Variance
  Standard deviation.



42

Characteristics of Variability

Frequency Distribution

Frequency:

Frequency refers to how often something occurs.

Array:

An array is an orderly arrangement of numbers, often in rows, columns or a matrix.

column	column	column	column	
12	23	4	-9	row
-7	13	6	0	row
9	-3	4	11	row

Array of Numbers
(3 rows x 4 columns)

43

Characteristics of Variability

Frequency Distribution

Frequency Distribution:

- **A frequency distribution is an arrangement of the values that one or more variables take in a sample, usually in table format.**
- **Each entry (class or category) in the table contains the frequency or count of the occurrences of values within a particular group or interval.**
- **The table summarizes the distribution of values in the sample.**

Scores:
1,1,2,2,2,2,2,3,3,3,3,4,4,5

Score	Frequency
1	2
2	5
3	4
4	2
5	1

44

Characteristics of Variability

Frequency Distribution

Example:

The following values are final marks in a test taken by 80 students:

68	84	75	82	68	90	62	88	76	93
73	79	88	73	60	93	71	59	85	75
61	65	75	87	74	62	95	78	63	72
66	78	82	75	94	77	69	74	68	60
96	78	89	61	75	95	60	79	83	71
79	62	67	97	78	85	76	65	71	75
65	80	73	57	88	78	62	76	53	74
86	67	73	81	72	63	76	75	85	77

45

Characteristics of Variability

Frequency Distribution

- **The highest mark is 97 and the lowest mark is 53.**
- **We can say the range is : 97 – 53 = 44.**
- **To determine the class interval: 44 / 10 = 4,4.**
- **Set the class interval as 5.**
- **Class intervals: [50,54], [55,59], [60,64], [65,69], ...**
- **[50,54] is a closed interval, contains values:**

  {50,51,52,53,54}

- **Class intervals do not overlap, accommodate all the data and are of the same size.**

46

Characteristics of Variability

Frequency Distribution

Class interval	Student marks (Mathematics)
50–54	53
55–59	59 57
60–64	62 60 61 62 63 60 61 60 62 62 63
65–69	68 68 65 66 69 68 67 65 65 67
70–74	73 73 71 74 72 74 71 71 73 74 73 72
75–79	75 76 79 75 75 78 78 75 77 78 75 79 79 78 76 75 78 76 76 75 77
80–84	84 82 82 83 80 81
85–89	88 88 85 87 89 85 88 86 75
90–94	90 93 93 94
95-99	95 96 95 97

47

Characteristics of Variability

Frequency Distribution

Re-arrange the marks in each class in ascending order (an ordered array):

Class interval	Student marks (Mathematics)
50–54	53
55–59	57 59
60–64	60 60 60 61 61 62 62 62 62 63 63
65–69	65 65 65 66 67 67 68 68 68 69
70–74	71 71 71 72 72 73 73 73 73 74 74 74
75–79	75 75 75 75 75 75 75 76 76 76 76 77 77 78 78 78 78 78 79 79 79
80–84	80 81 82 82 83 84
85–89	85 85 85 86 87 88 88 88 89
90–94	90 93 93 94
95-99	95 95 96 97

48

Characteristics of Variability

Frequency Distribution

49

Characteristics of Variability

Frequency Distribution

Conclusions out of the frequency distribution:

- **The highest mark is 97**
- **The lowest mark is 53**
- **The range is 97 – 53 = 44**
- **The five highest ranking students have marks {97, 96, 95, 95, 94}**
- **The five lowest ranking students have marks {53, 57, 59, 60, 60}**
- **The mark of the student ranking tenth highest is 88**
- **The number of students receiving marks of 75 or higher is 44**
- **The number of students receiving marks below 85 is 63**
- **The marks which did not appear are**

{0–52, 54-56, 58, 64, 70, 91, 92, 98-100}

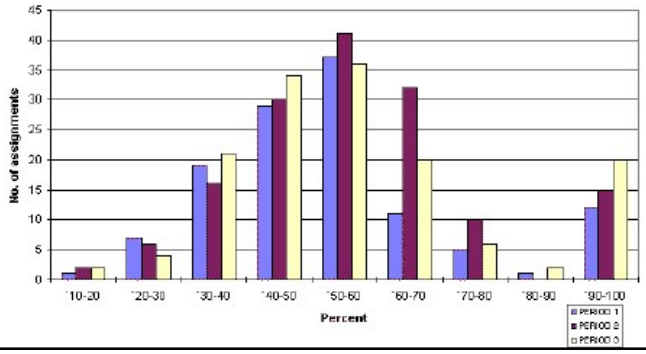
50

Characteristics of Variability

Histogram

Frequency Histogram:

- **Defined as a graphical representation of data.**
- **The data is grouped into ranges (such as "40 to 49"), and then plotted as bars.**
- **Similar to a Bar Graph, but each bar represents a range of data.**



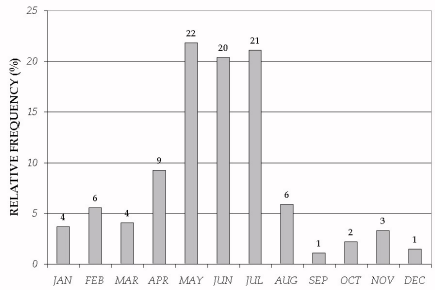
51

Characteristics of Variability

Histogram

Relative Frequency Histogram:

- **Shows the proportion of events rather than the absolute number of events.**
- **The relative frequency of a class is the frequency of the class divided by the total number of frequencies of the class and is generally expresses as a percentage.**
- **The sum of the relative frequencies of all classes is 1 or 100 %.**



52

Characteristics of Variability

Histogram

Cumulative Frequency Distribution:

- The total of a frequency and all frequencies below it in a frequency distribution.
- It is the 'running total' of frequencies.

Example:

Consider the following table populated with the weight of a hundred students:

Mass (kg)	Frequency (number of students)
60-62	5
63-65	18
66-68	42
69-71	27
72-74	8
	Total 100

53

Characteristics of Variability

Histogram

Example (Continued):

- The cumulative frequency up to and including the class interval 66-68 is:

$$5 + 18 + 42 = 65$$

- 65 students have masses less than 68,5 kg

Cumulative Frequency Table of previous data:

Mass (kg)	Frequency (number of students)
less than 59,5	0
less than 62,5	5
less than 65,5	23
less than 68,5	65
less than 71,5	92
less than 74,5	100

54

Characteristics of Variability

Histogram

Example:

The following data set contains 80 compressive strengths of aluminium-lithium alloy specimens:

105	221	183	186	121	181	180	143
97	154	153	174	120	168	167	141
245	228	174	199	181	158	176	110
163	131	154	115	160	208	158	133
207	180	190	193	194	133	156	123
134	178	76	167	184	135	229	146
218	157	101	171	165	172	158	169
199	151	142	163	145	171	148	158
160	175	149	87	160	237	150	135
196	201	200	176	150	170	118	149

55

Characteristics of Variability

Histogram

Example (Continued):

- **The following distribution table contains nine classes, each of width 20 units:**

Class interval (psi)	Tally	Observed Frequency	Relative frequency	Cumulative relative frequency
70 – 89	||	2	0,0250	0,0250
90 – 109	|||	3	0,0375	0,0625
110 – 129	~~||||~~ |	6	0,0750	0,1375
130 - 149	~~||||~~ ~~||||~~ ||||	14	0,1750	0,3125
150 - 169	~~||||~~ ~~||||~~ ~~||||~~ ~~||||~~ ||	22	0,2750	0,5875
170 - 189	~~||||~~ ~~||||~~ ~~||||~~ ||	17	0,2125	0,8000
190 - 209	~~||||~~ ~~||||~~	10	0,1250	0,9250
210 - 229	||||	4	0,0500	0,9750
230 - 249	||	2	0,0250	1,0000

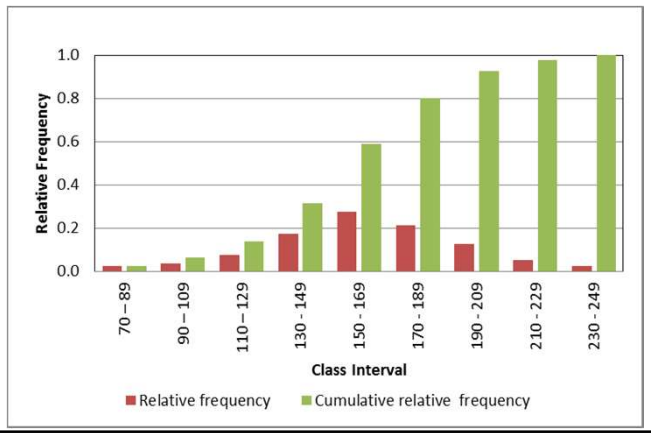
56

Characteristics of Variability

Histogram

Example (Continued):

- **The following depicts the relative and cumulative relative frequency of the previous data:**



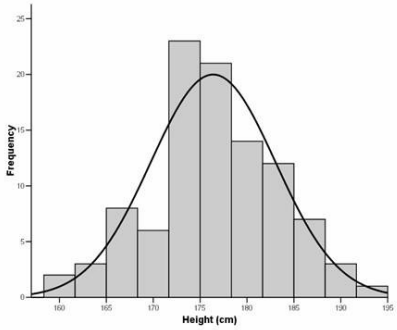
57

Characteristics of Variability

Normal Distribution

Normal distribution:

- **Follows a "Bell" curve.**
- **The centre contains the greatest number of a value (Mean) and therefore is the highest point on the arc of the line.**
- **The curve is concentrated in the centre and decreases on either side.**



58

Characteristics of Variability

Normal Distribution

Normal distribution (Continued):

- **The bell curve signifies that the data is symmetrical and thus we can create reasonable expectations as to the possibility that an outcome will lie within a range to the left or right of the centre, once we can measure the amount of deviation (σ) contained in the data.**
- **The mean (μ) identifies the position of the centre and the standard deviation (σ) determines the height and width of the bell.**
- **A large standard deviation creates a bell that is short and wide while a small standard deviation creates a tall and narrow curve.**
- **The mean and the median are the same in a normal distribution.**
- **The area under the curve is equal to one.**

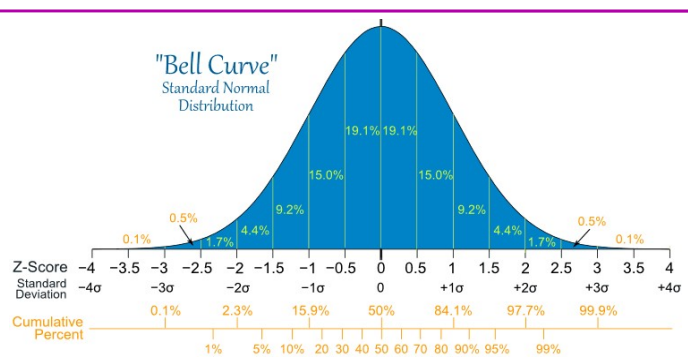
59

Characteristics of Variability

Normal Distribution

Normal distribution (Continued):

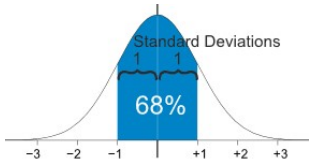
- **The total area under the curve is equal to 1 (100%).**
- **About 68% of the area under the curve falls within 1 standard deviation.**
- **About 95% of the area under the curve falls within 2 standard deviations.**
- **About 99.7% of the area under the curve falls within 3 standard deviations.**



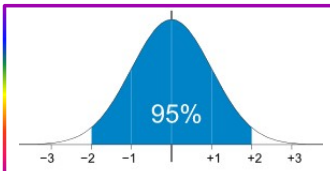
60

Characteristics of Variability

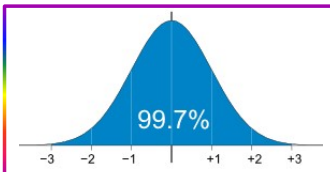
Normal Distribution



68% of values are within 1 standard deviation of the mean



95% of values are within 2 standard deviations



99.7% of values are within 3 standard deviations

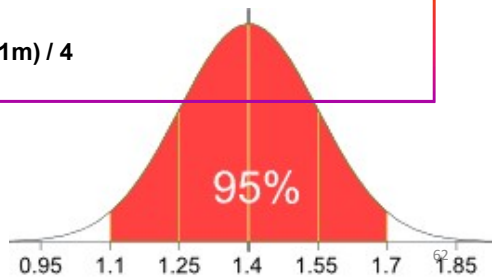
61

Characteristics of Variability

Normal Distribution

Example:

- **95% of students at school are between 1.1m and 1.7m tall.**
- **Assuming this data is normally distributed the mean and standard deviation can be calculated:**
- **The mean is halfway between 1.1m and 1.7m:**
- **Mean = (1.1m + 1.7m) / 2 = 1.4m**
- **95% is 2 standard deviations either side of the mean (a total of 4 standard deviations) so:**
- **1 standard deviation = (1.7m-1.1m) / 4**
  = 0.6m / 4 = 0.15m



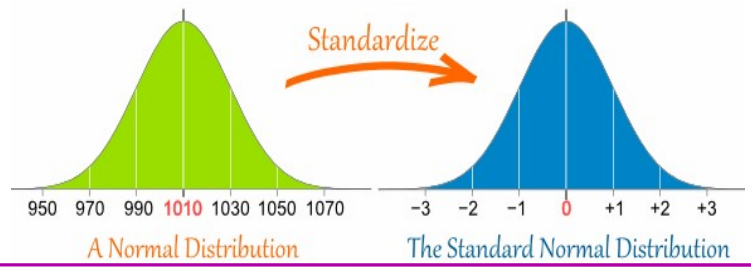
62

Characteristics of Variability

Normal Distribution

Standard score:

- The number of standard deviations from the mean is also called the "Standard Score", "sigma" or "z-score".
- How to convert a value to a Standard Score ("z-score"):
  - first subtract the mean,
  - then divide by the Standard Deviation
- And doing that is called "Standardizing":



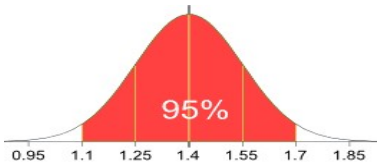
63

Characteristics of Variability

Normal Distribution

Example:

In the previous example in that same school one of your friends is 1.85m tall.



- Out of the bell curve it can be seen that 1.85m is 3 standard deviations from the mean of 1.4, so:
  - Your friend's height has a "z-score" of 3.0
- It is also possible to calculate how many standard deviations 1.85 is from the mean.
- How far is 1.85 from the mean?
  - It is 1.85 - 1.4 = 0.45m from the mean
- How many standard deviations is that?
- The standard deviation is 0.15m, so:
  - 0.45m / 0.15m = 3 standard deviations

64

Characteristics of Variability

Normal Distribution

Example:

- How to use table 8 in the notes:
- What is the Percentage of Population Between 0 and 0.45?
- In table 8 is the "0.1"s running down, and then the "0.01"s running along (to the right).
- Start at the row for 0.4, and read along until 0.45: there is the value 0.1736, this is the area under the curve between 0 and 0.45.
- So 0.1736 of the population are between 0 and 0.45 Standard Deviations from the Mean.
- And 0.1736 is 17.36%.
- So 17.36% of the population are between 0 and 0.45 Standard Deviations from the Mean.

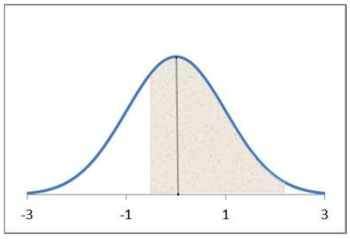
65

Characteristics of Variability

Normal Distribution

Example:

- Find the area under the normal curve between z = -0,46 and z = 2,21



- Area = Area [z: -0,46 0] + [z: 0 2,21]
  = Area [z: 0,46 0] + [z: 0 2,21}
  = 0,1772 + 0,4864
  = 0,6636

- The result 0,6636 is the required area and represents the probability that z is between -0,46 and 2,21 , denoted by:

$$Pr\{ -0,46 \leq z \leq 2,21\}$$

66

Characteristics of Variability

Standard Error:

- Represents how well the sample mean approximates the population mean, i.e. the standard error gives a measure of how well a sample represents the population.
- The larger the sample, the smaller the standard error, and the closer the sample mean approximates the population mean.
- When the sample is representative, the standard error will be small.
- The standard error is the standard deviation divided by the square root of N (the sample size):

$$\text{Standard Error } (s_{\bar{x}}) = \frac{s}{\sqrt{n}}$$

where:

S = Standard Deviation (sample)
n = quantity of numbers in group

67

Characteristics of Variability

Example:

- Consider the following sample sets:

	Sample 1	Sample 2	Sample 3	Sample 4
	9	6	5	8
	2	6	3	1
	1	8	6	7
		8	4	1
		3	7	3
		8	2	3
			6	4
			9	7
			7	1
			1	8
			1	9
			7	9
				3
				1
				6
				8
				3
				4
Mean:	4	6.5	4.83	4.78
Std dev, s:	4.36	1.97	2.62	2.96
Sample size, n:	3	6	12	18
sqrt(n):	1.73	2.45	3.46	4.24
Standard error, s/sqrt(n):	2.52	0.81	0.76	0.70

- Note how the standard error reduces with increasing sample size

68

Characteristics of Variability

Confidence Level:

- The confidence level tells you how sure you can be .
- Confidence level is a percentage, typically 90%, 95% or 99%.
- The higher the level of confidence, the wider the interval.
- The lower the level of confidence, the shorter the interval.
- A confidence level of 95% means that 19 out of 20 times the confidence interval around the population mean μ is right.

For a confidence level of 95%:

Confidence interval:

$$\mu = \bar{x} \pm z_{0,025} \frac{\sigma}{\sqrt{n}}$$

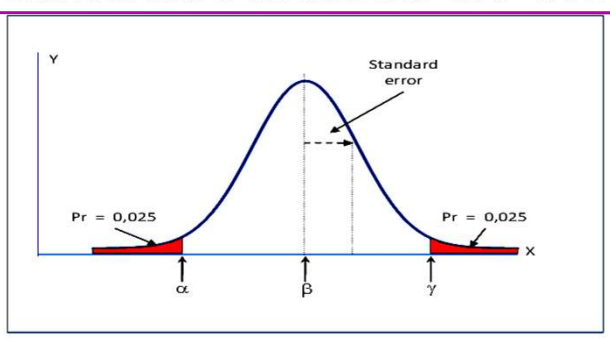
69

Characteristics of Variability

Confidence Interval:

A confidence interval is a range around a measurement that conveys how precise the measurement is.

- For a confidence level of 95%, select the smallest range under the normal distribution of $\bar{x}$ that will enclose a probability of 95%.
- This leaves 2,5% probability excluded under half the normal distribution.
- For this we need a z-value of 1,96 (Table 8: 0,5 -0,025 = 0,475)



70

Characteristics of Variability

Example:

Sixteen marks were sampled randomly from a very large class that had a standard deviation of 12; these 16 marks had a mean of 58. Find a 95% confidence interval for the mean mark of the whole class.

$$n = 16$$
$$\sigma = 12$$
$$\bar{x} = 58$$

$$\mu = \bar{x} \pm z_{0,025} \frac{\sigma}{\sqrt{n}}$$

$$= 58 \pm (1,96)\frac{12}{\sqrt{16}}$$

$$= 58 \pm 6$$

The 95% confidence interval for μ is:

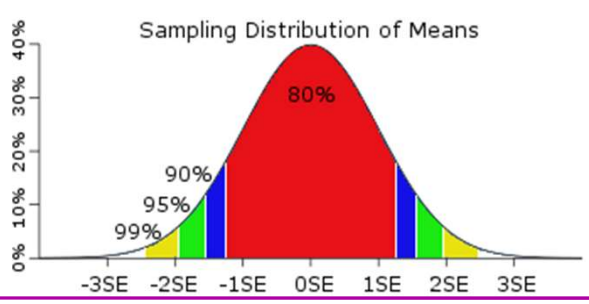
$$52 < \mu < 64$$

71

Characteristics of Variability

For other confidence intervals:

Confidence level	Standard Errors from mean
99%	2.58
95%	1.96
90%	1.64
80%	1.28



72

Statistical Hypothesis

Hypothesis:

- A hypothesis is an explanation for a set of observations.
- Most hypothesis are either "If, then" statements or else forms of the null hypothesis.

Null Hypothesis:

- A proposition that undergoes verification to determine if it should be accepted or rejected in favour of an alternative proposition.
- Often the null hypothesis is expressed as "There is no relationship between two quantities."

Example:

- In market research, a null hypothesis would be "A ten-percent increase in price will not adversely affect the sale of this product."
- Based on survey results, this hypothesis will be either accepted as correct or rejected as incorrect.

73

Statistical Hypothesis

Type I Error:

- False positive.
- Usually a type I error leads one to conclude that a thing or relationship exists when really it doesn't.

Example of Type I Error:

A patient has a disease for which he is being tested when really the patient does not have the disease.

Type II Error:

- False negative.
- Usually a type II error leads one to conclude that a thing or relationship which is perceived as not existing, actually exists.

Example of Type II Error:

A blood test failed to detect the disease it was designed to detect, in a patient who really has the disease.

74

Statistical Hypothesis

Measured / Perceived	Reality: True	Reality: False
True	Correct ☺	Type I False Positive
False	Type II False Negative	Correct ☺

75

Central Limit Theory

Central Limit Theory:

- **The Central Limit Theorem describes the characteristics of the "population of the means", which has been created from the means of an infinite number of random population samples of size (N), all of them drawn from a given "parent population".**

- **i.e. :**
  Let the set of sample means be

$$\bar{X} = \{\bar{x}_1, \bar{x}_2, \bar{x}_3, \ldots, \bar{x}_j, \ldots, \bar{x}_\infty\}$$

and the set of sample standard deviations be

$$S = \{s_1, s_2, s_3, \ldots, s_j, \ldots, s_\infty\}$$

- ***The Central Limit Theory states regardless of the distribution of the population, the distribution of $\bar{X}$ will tend to follow a normal curve as the sample size increases (>30).***

76

Tests of Significance

t – distribution:

- The t-distribution is generally used to estimate the uncertainty about the population mean μ.
- Sample size is usually small.
- It is symmetrical, bell-shaped, and similar to the standard normal curve.
- It differs from the standard normal curve, however, in that it has an additional parameter, called degrees of freedom (*v*), which changes its shape.

If $\bar{x}$ is the mean of a random sample of size n taken from a normal population having the mean μ and the variance σ^2, then

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

With $v = n - 1$

77

Tests of Significance

Example:

A manufacturer of fuses claims that with a 20% overload, the fuses will blow in 12,40 minutes on the average.
To test this claim, a sample of 20 fuses was subjected to a 20% overload, and the times it took them to blow had a mean of 10,63 minutes and a standard deviation of 2,48 minutes.

If it can be assumed that the data constitute a random sample from a normal population, do they tend to support or refute the manufacturer's claim?

$$t = \frac{|\bar{x} - \mu|}{s/\sqrt{n}} = \frac{|10{,}63 - 12{,}40|}{2{,}48/\sqrt{20}} = 3{,}19$$

- $v = 20 - 1 = 19$
- Out of table 9: probability that t will exceed 2,86 is 0,005
- t = 3,19 is more than 2,86; 0,005 is a very small probability.
- In all probability the mean blowing time of the fuses with a 20% overload is less than 12,40 minutes. (difference is significant)

78

Tests of Significance

Example:

Consider the following results on ten batches of asphalt to determine the bitumen content, as tested by two laboratories:

Batch no.	1	2	3	4	5	6	7	8	9	10
Lab A	5.5	6.6	5.7	4.2	6.8	5.2	7.2	4.7	5.9	5.3
Lab B	6.1	6.6	5.8	4.8	6.5	5.9	7.2	5.0	6.0	5.2

Conduct the analysis in terms of the discrepancies between the laboratories:

Batch no.	1	2	3	4	5	6	7	8	9	10
Discrepancy	0.6	0.0	0.1	0.6	-0.3	0.7	0.0	0.3	0.1	-0.1

79

Tests of Significance

Example (Continued):

The mean value of the discrepancies is:

$$\bar{x} = \frac{0{,}6+0{,}0+0{,}1+0{,}6+(-0{,}3)+0{,}7+0{,}0+0{,}3+0{,}1+(-0{,}1)}{10}$$

$$= \frac{2}{10}$$

$$= 0{,}20$$

The standard deviation is: square root (square of discrepancy – mean)

$$s = \sqrt{\frac{0{,}16+0{,}04+0{,}01+0{,}16+0{,}25+0{,}25+0{,}04+0{,}01+0{,}01+0{,}09}{10-1}}$$

$$= 0{,}35$$

80

Tests of Significance

Example (Continued):

$$t = \frac{|0{,}20 - 0|}{0{,}35/\sqrt{10}}$$

$$= 1{,}81$$

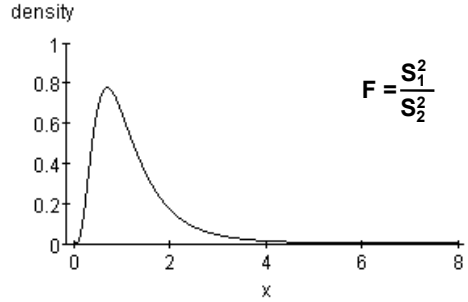
- **Degree of freedom = v = 10 – 1 = 9**
- **From table 9: I $t_{9;0,025}$ I = 2,26**
- **The calculated value of *t* is less than the value that would be exceeded by chance only 5% of the time.**
- **The difference between laboratories is not significant.**

81

Tests of Significance

F-distribution:

- **t-test is used to compare two means, F-test is based on the ratio between two variances, i.e. compare the variance of two samples for significance.**
- **F-values are all positive.**
- **The F-distribution is not symmetrical.**
- **The F-distribution is skewed to the right and has a minimum value of 0, but no maximum value.**
- **F values depend on 2 parameters: numerator degrees of freedom (v_1) and denominator degrees of freedom (v_2).**



$$F = \frac{S_1^2}{S_2^2}$$

82

Tests of Significance

Example:

- Laboratory A and Laboratory B are required to test 25 concrete cubes representative of the same batch at a certain date.
- The purpose of the exercise was to check whether the variability that could be assigned to the test results is different at a 5 % level of significance.

- Data obtained from the laboratories:

Laboratory	A	B
mean	32,0	36,2
S.D.	3,35	2,43
n	25	25

- 5% level of significance: table 10(a) for $f_{0,05\ v1\ v2}$
- Degrees of freedom: v1 = v2 = 25 – 1= 24
- $S_A^2 = 3,35^2$
- $S_B^2 = 2,43^2$

83

Tests of Significance

Example (Continued):

From table 10(a): $F_{0,05\ 24;24} = 1,98$

and

$$F = \frac{S_A^2}{S_B^2} = \frac{3,35^2}{2,43^2} = 1,90$$

$F < F_{0,05\ 24;24}$ which means that the F value is not significant, the two variances are equal.

84

Tests of Significance

Chi-square distribution:

- Chi-square (χ^2)is a statistical test commonly used to compare observed data with data we would expect to obtain according to a specific hypothesis.
- The chi-square test is always testing what scientists call the null hypothesis, which states that there is no significant difference between the expected and observed result.
- Chi-square requires that you use numerical values, not percentages or ratios.
- The chi-square is the sum of the squared difference between observed (*o*) and the expected (*e*) data, divided by the expected data in all possible categories:

$$\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$$

Where:
O represents the observed frequency
E represents the expected frequency

85

Tests of Significance

Example:

In 200 tosses of a coin, 115 heads and 85 tails were observed. We are required to test the hypothesis that the coin is fair.

- Observed frequencies are:

Heads: $O_1 = 115$
Tails: $O_2 = 85$

- Expected frequencies are:

Heads: $E_1 = 100$
Tails: $E_2 = 100$

- Then $\chi^2 = \sum_i \frac{(O_i - E_i)^2}{E_i}$

$$= \frac{(115-100)^2}{100} + \frac{(85-100)^2}{100}$$

$=4,50$

86

Tests of Significance

Example (Continued):

- **Number of categories or classes is 2**
- **Hence degrees of freedom, *v*, is 1**
- **Out of Table 12: $\chi^2_{.95}$ = 3,84**
- **Since 4,50 > 3,84**

We reject the hypothesis that the coin is fair at the 95% level of significance.

- **However, at the critical value of $\chi^2_{.99}$ for *v* = 1 is 6,63 and since**

4,50 < 6,63

we cannot reject the hypothesis that the coin is fair at this level of confidence.

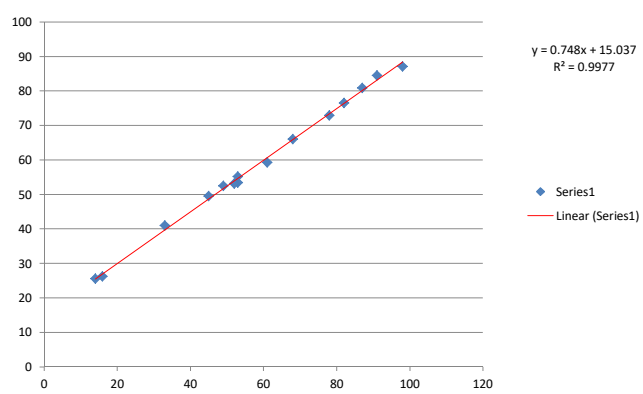
87

Regression Analysis

- **Regression analysis is a statistical tool for the investigation of relationships between variables.**
- **Regression analysis is widely used for prediction and forecasting**

Example:

x	y
53	53.42
91	84.53
87	80.87
49	52.45
14	25.6
98	87.12
78	72.93
82	76.5
53	55.14
33	41
45	49.51
52	53.16
16	26.22
68	66.04
61	59.26

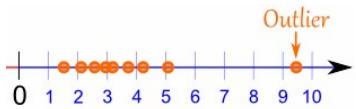


88

Outliers

Outliers:

- Outliers are extreme cases which can strongly influence statistical calculations.
- An outlier can be a value that "lies outside" (is much smaller or larger than) most of the other values in a set of data.



- To determine whether a suspicious value is an outlier, the status of the standard deviation available has to be known.

$$C_o = |x_o - \bar{x}|$$

- Where: x_o = the suspicious value in a set of observations
  $\bar{x}$ = mean of the set of observations.

- If $C_o > C_{0,01;\,n}\,\sigma$ (see Table 13)

the suspicious value x_o, is regarded as an outlier and should be excluded from the set of data.

89

Outliers

Example:

- Let the dataset

{X: 5,1 5,2 5,2 5,2 5,0 5,6}

represent percentage bitumen contents. The mean is 5,22 and the sample standard deviation s is 0,20. Determine whether the suspicious value 5,6 is an outlier.

- The population standard deviation for percentage bitumen content is 0,3% (table 11(a)).

$\bar{x}$ = 5,2

x_o = 5,6

σ = 0,3

$$C_o = |x_o - \bar{x}| = |5{,}6 - 5{,}2| = 0{,}4$$

- and

$C_{0,01;\,6}\,\sigma$ = (2,68)(0,3) = 0,8 (Table 13, n = 6)

- hence $C_o < C_{0,01;\,6}$
- and the suspicious value x_o is NOT an outlier.

90

Uncertainty of Measurement

Measurement:

- A measurement tells us about a property of something.
- A measurement tells us how heavy an object is, or how hot, or how long it is.
- A measurement gives a number to that property.
- The result of a measurement is normally in two parts: a number and a unit of measurement, e.g. "How long is it? ... 2 metres."

Uncertainty of Measurement:

- Uncertainty of measurement is the doubt that exists about the result of any measurement.
- For every measurement - even the most careful - there is always a margin of doubt.
- This might be expressed as "give or take" ... e.g. a stick might be two metres long "give or take a centimetre".
- In order to quantify an uncertainty the width of the margin, or interval, and the confidence level (which states how sure we are that the "true value" is within that margin) are needed.

91

Uncertainty of Measurement

Example:

The measurement is: 5,07 g ± 0,02 g.

This implies that the experimenter is confident that the actual value for the quantity being measured lies between 5,05 g and 5,09 g.

Always round the experimental measurement or result to the same decimal place as the uncertainty.

Wrong: 52,3 cm ± 4,15 cm
Correct: 52 cm ± 4 cm

92

Uncertainty of Measurement

- If the ranges of two measured values don't overlap, the measurements are *discrepant* (the two numbers don't agree).
- If the ranges overlap, the measurements are said to be *consistent*

Measurements don't agree	0,86 s ± 0,02 s and 0,98 s ± 0,02 s
Measurements agree	0,86 s ± 0,08 s and 0,98 s ± 0,08 s

Uncertainty values associated with a number of tests on road building materials are given in Table 11 (a) to (c)

93

Uncertainty of Measurement

Error:

- Error is the difference between the measured value and the 'true value' of the thing being measured.
- Error can be dealt with, as it is usually the result of something the person who is conducting the experiment is doing or as a result of something that is happening with the equipment.
- Whenever possible, one strives to correct for any known errors. One can do so, e.g., by applying corrections from calibration certificates.

Uncertainty:

- Uncertainty is a quantification of the doubt about the measurement result.
- Uncertainty is something that is inherent in the experiment and cannot be eliminated., it can only be characterised.
- Any error whose value we do not know is a source of uncertainty.
- Uncertainty cannot be avoided but it can be reduced by using 'better' apparatus.

94

Uncertainty of Measurement

Repeatability:

The difference between two test results obtained by the same operator with the same apparatus under constant operating conditions on identical test material.

Reproducibility:

The difference between two single and independent test results obtained by different operators working in different laboratories on identical test material

95

Please bring a Calculator for the workshop Avoid HP programmable calculators, bring a Casio or Sharp scientific calculator

96

End of Chapter 2

97