5.1 PRESENTATION OF DATA

DISCRETE DATA ⇒ Exact number values. Can be counted. E.g. Number of people, test scores, shoe sizes, ...

CONTINUOUS DATA ⇒ Takes numerical values within a range, usually measured. E.g. Height, weight, temperature, ...

PRESENTATION METHODS:

TABLES ⇒ We start with a list:

- This data can be shown with a frequency table (right), and a grouped frequency table w/ equal class sizes:

<table>
<thead>
<tr>
<th>SCORE</th>
<th>TALLY</th>
<th>FREQ.</th>
<th>RELATIVE FREQ.%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1-2</td>
<td>III</td>
<td>3</td>
<td>10%</td>
</tr>
<tr>
<td>3-4</td>
<td>II</td>
<td>2</td>
<td>6.7%</td>
</tr>
<tr>
<td>5-6</td>
<td>IIIII</td>
<td>5</td>
<td>16.7%</td>
</tr>
<tr>
<td>7-8</td>
<td>IIIII II</td>
<td>12</td>
<td>40%</td>
</tr>
<tr>
<td>9-10</td>
<td>IIIII</td>
<td>8</td>
<td>26.7%</td>
</tr>
<tr>
<td><strong>TOTAL:</strong> 30</td>
<td>100%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- The next step here would be to display the data in a HISTOGRAM ⇒

We can say here that the 'modal class' is a score of 7 to 8.

BOX & WHISKER PLOTS ⇒ This is another way of showing your data. It shows the median, upper/lower quartiles (see 5.2) and 'outliers'.

→ Outliers: These are data points that lie further than 1.5 × IQR from the median.
5.2 - STATISTICAL MEASURES

CENTRAL TENDENCY:

MODE - Most frequent value/class:

MEDIAN - Order data, then observe the \( \frac{(n+1)}{2} \)th value:

MEAN - \( \frac{\sum x}{\sum f} \):

For grouped frequency:

You can only 'estimate' the total by multiplying midpoints by freq. and summing those.

\[ \frac{\sum (f \cdot x)}{\sum f} \]

QUARTILES - Similar to how the median splits the data into 50% parts, quartiles are 4 equal parts of 25%, after ordering.

LOWER QUARTILE (LQ/Q1) - Point that is greater than 25% of the data

UPPER QUARTILE (UQ/Q3) - Point that is greater than 75% of the data

PERCENTILES - An extension of this breakdown of the data is splitting it into 1% sections, called 'percentiles'.

DISPERSION

RANGE - HIGHEST VALUE - LOWEST VALUE = RANGE

INTERQUARTILE RANGE - \( IQR = Q_3 - Q_1 \)

VARIANCE - Another measure of spread of data. It is essentially the average of the differences between each value and the mean:

\[ (\text{Var. } = ) S_n^2 = \frac{\sum (x_i - \bar{x})^2}{n} \]

STANDARD DEVIATION - Related measure, just calculated by doing the square root of the variance. S.D. is often used more frequently.

CALCULATOR:

\( \text{TI-nspire: Enter values in 'Lists & Spreadsheets', then on a calculator page, press MENU \rightarrow 6: Statistics} \rightarrow 1: \rightarrow 1: \text{One-var stats. S.D. is } \sigma \).

\( \text{TI-84: Press STAT} \rightarrow 1:\text{EDIT} \rightarrow \text{ENTER}, \text{ then fill column with values. Then press STAT} \rightarrow \text{CALC} \rightarrow 1:1-\text{VAR STATS} \rightarrow \text{ENTER. S.D. is } \sigma \).
5.3 - CUMULATIVE FREQUENCY

It is hard to see what proportion of the data is above or below a certain value. Adding on the frequency of each successive class, we build a 'cumulative freq.' column. We plot this against the UB of each class, then analyse:

### Example

<table>
<thead>
<tr>
<th>SCORE</th>
<th>FREQ.</th>
<th>U.B.</th>
<th>C.F.</th>
</tr>
</thead>
<tbody>
<tr>
<td>10≤x&lt;20</td>
<td>2</td>
<td>20</td>
<td>2</td>
</tr>
<tr>
<td>20≤x&lt;30</td>
<td>5</td>
<td>30</td>
<td>7</td>
</tr>
<tr>
<td>30≤x&lt;40</td>
<td>7</td>
<td>40</td>
<td>14</td>
</tr>
<tr>
<td>40≤x&lt;50</td>
<td>21</td>
<td>50</td>
<td>35</td>
</tr>
<tr>
<td>50≤x&lt;60</td>
<td>36</td>
<td>60</td>
<td>71</td>
</tr>
<tr>
<td>60≤x&lt;70</td>
<td>40</td>
<td>70</td>
<td>111</td>
</tr>
<tr>
<td>70≤x&lt;80</td>
<td>27</td>
<td>80</td>
<td>138</td>
</tr>
<tr>
<td>80≤x&lt;90</td>
<td>9</td>
<td>90</td>
<td>147</td>
</tr>
<tr>
<td>90≤x&lt;100</td>
<td>3</td>
<td>100</td>
<td>150</td>
</tr>
</tbody>
</table>

### Analysis

#### Finding Median

- Start at the \(\left(\frac{n}{2}\right)^{th}\) position on y-axis, trace across to the curve, then down to the x-axis, then read your median. (See blue line)

#### Lower Quartile

- Start at the \(\left(\frac{3(n+1)}{4}\right)^{th}\) position, then follow the same process as above.

#### Upper Quartile

- Start at the \(\left(\frac{3(n+1)}{4}\right)^{th}\) position, then follow the same process as above.

#### Finding Percentiles

- Similar process, but with \(\left(\frac{p(n+1)}{100}\right)^{th}\) if you are trying to find the \(p^{th}\) percentile.

\[ \text{E.G. (above): } 90^{th} \text{ percentile } : \frac{90(151)}{100} = 135.9 \]

\[ \text{E.G. (above): } \text{Median} = 61 \]

\[ \text{LQ} = 51 \]

\[ \text{IQR} = 71 - 51 = 20 \]

\[ \text{UQ} = 71 \]

\[ 90^{th} \text{ percentile} = 79 \]
Correlation

Testing the correlation between two sets of data, \( x \) and \( y \), means that you are testing whether a change in \( x \) causes a similar change in \( y \), and to what extent.

Graphs ⇒ This is one way of showing what the strengths of correlation mean in reality:

- **No correlation**
- **Weak positive correlation**
- **Strong positive correlation**
- **Very strong negative correlation**

Correlation Coefficient ⇒ An accurate method of calculating correlation, \( r \), measured on a scale from -1 (perfect negative corr.) to +1 (perfect positive correlation).

By Hand: \( r = \frac{\sum(x-x)(y-y)}{\sqrt{\sum(x-x)^2 \sum(y-y)^2}} \)

(You will use a calculator in exams)

By Calc:

**TI - 84**

- **STAT** \( 1: \) Edit \( \rightarrow \) ENTER
- Put \( x \) values in L1, \( y \) in L2
- **MODE** Check diagnostics are on
- **2nd** \( \text{QUIT} \)
- **STAT** \( 4: \) LinReg \( (ax+b) \) \( \rightarrow \) ENTER until info shown

**TI - Nspire**

- Add 'Lists & Spreadsheet'
- Name columns: 'x' & 'y'
- Type data
- **MENU** \( 4: \) Stats \( 1: \) Stat Calc \( 3: \) Lin Reg \( (mx+b) \)
- Choose 'x' for X list, 'y' for Y list
- **ENTER**, then observe

Equation of Regression Line ⇒ On the same calculator screen as the \( r \) value, you will see a value for \( a \) & \( b \). This creates an equation in the form \( y=ax+b \).

This is a more advanced version of what you may have seen as a 'line of best fit', which is more of an estimate.

You may well be asked to use this equation to estimate \( y \) values given \( x \) values.
5.5 - PROBABILITY

DEFINITIONS
'TRIAL' → Each time an experiment is repeated.
'OUTCOMES' → Possible results of one trial.
'SAMPLE SPACE (U)' → Set of all possible outcomes in an experiment.

PROB RULES
If A is a set of results from an experiment with all equally likely results, then:

\[ \text{Prob. of } A \text{ occurring} = P(A) = \frac{\text{EVENTS IN A}}{\text{EVENTS IN U}} = \frac{n(A)}{n(U)} \]

E.g. → Probability of rolling a multiple of 3 on a fair dice:

\[ \text{sol. } P(\text{Mult. of 3}) = \frac{2}{6} = \frac{1}{3} \]

→ Two events are complementary if exactly one of the two events must occur, so:

\[ P(A) + P(A') = 1 \]

E.g. → Rolling a 6 and not rolling 6 are complementary, as the probabilities add to 1:

\[ \text{sol. } \frac{1}{6} + \frac{5}{6} = 1 \]

VENN DIAGRAMS
The venn diagram on the left represents the earlier multiples of 3 problem. You can see where the 2 out of 6 probability is derived from:

INTERSECTION
The venn to the right represent the 'intersection' between sets A & B, denoted \(A\cap B\). These are the elements common to both A and B:

UNION
Denoted \(A\cup B\), represents elements in A or B or both:
PROBABILITY TREES

When you have two or more trials, and the possible outcomes are not too numerous, we can use 'probability trees':

E.g.: In archery, Jim hits a target \( \frac{2}{5} \) of the time, and Lucy hits it \( \frac{4}{5} \) of the time.

i) Construct tree:

\[
\begin{align*}
& \text{Jim} \\
\frac{2}{5} & \quad \frac{3}{5} \\
H & \quad M
\end{align*}
\]

\[
\begin{align*}
\text{LUCY} & \quad \text{OUTCOME} & \quad \text{PROB.} \\
H & \quad H & \quad \frac{2}{5} \times \frac{4}{5} = \frac{8}{20} \\
M & \quad M & \quad \frac{2}{5} \times \frac{4}{5} = \frac{8}{20}
\end{align*}
\]

ii) Prob. of at least one hit?

\[
\frac{1}{20} + \frac{4}{20} + \frac{12}{20} = \frac{17}{20}
\]

5.6 - COMBINED PROB.

UNION PROB. \( P(A \cup B) = P(A) + P(B) - P(A \cap B) \)

INTERSECTION \( P(A \cap B) = P(A) \times P(B) \) [if independent]

CONDITIONAL \( P(A \mid B) = \frac{P(A \cap B)}{P(B)} \) → where \( P(A \mid B) \) means the prob. of \( A \) occurring, given that \( B \) has occurred.

INDEPENDENT → Two events are independent if one event occurring does not affect the probability of the other event occurring.

It can also be shown formally by checking whether:

\( P(A \mid B) = P(A) \) holds,  

or

\( P(A \cap B) = P(A) \times P(B) \) holds.
5.7 - DISCRETE RANDOM VARIABLES

**Random Variable** ⇒ this represents, in number form, the possible outcomes which could occur for some random experiment.

**Discrete** ⇒ A set of distinct possible values, which you can count.

**Continuous** ⇒ Values are measured between a certain range.

**Probability Distributions** ⇒ For any random variable, there is a prob. dist. which describes the probability of each value occurring.

- Notation: The prob. that the variable $X$ takes value $x$ is denoted: $P(X=x)$.
- E.g.: Tossing 2 coins, counting how many 'heads' occur.
  - $P(X=0) = \frac{1}{4}$, $P(X=1) = \frac{2}{4} = \frac{1}{2}$, $P(X=2) = \frac{1}{4}$

**Rule** ⇒ For something to be a valid probability distribution function, the sum of the probabilities needs to equal 1.

- I.e.: $\sum P(x_i) = 1$  
  - E.g. ⇒ (From above) $\frac{1}{4} + \frac{1}{2} + \frac{1}{4} = 1$
  - E.g. ⇒ Find $k$: $\frac{x}{P(x=x)} = \begin{array}{c|c} 0 & 1 \end{array} \frac{1}{0.3 \ 0.5} \rightarrow 1 - 0.3 - 0.5 = 0.2 \ 0.2 \ k = 0.2$

**Expected Value** ⇒ Take $n$ trials of an experiment, where in each of the trials the event has prob. of $p$ of occurring, then the number of times we expect the event to occur is $n \times p$.

The expected outcome for the random variable $X$ is the mean result, $\mu$:

- E.g. ⇒ In a magazine store, 23% of customers purchased 1 magazine, 38% bought 2, 21% bought 3, 13% bought 4, and 5% bought 5. Calculate the expected number of magazines bought:
  - Probability table:
    - $x$: 1 2 3 4 5
    - $p_i$: 0.23 0.38 0.21 0.13 0.05
    - $\mu = \sum x_i p_i$
    - $= 1(0.23) + 2(0.38) + 3(0.21) + 4(0.13) + 5(0.05)$
    - $= 2.39$ magazines
**Binomial Experiments**

These are experiments where there are just 2 possible results: success or failure (event occurring or not). This is then repeated in a number of independent trials. For each trial, prob. of success is $p$, and prob. of failure is therefore $1-p$.

**Opening Problem:** A 'spinner' has three blue edges and one white edge so it has $\frac{3}{4}$ prob. of getting blue, and $\frac{1}{4}$ of getting white.

If we call spinning a blue a 'success', so $p = \frac{3}{4}$, then we can analyse the prob. of getting $0, 1, 2 \& 3$ successes from 3 spins.

**Probabilities**

\[
\begin{align*}
P(3 \text{ blues}) &= P(X = 3) = \left(\frac{3}{4}\right)^3 = 0.4219 \\
P(0 \text{ blues}) &= P(X = 0) = \left(\frac{1}{4}\right)^3 = 0.0156 \\
\end{align*}
\]

Those two are easy to calculate, but 1 blue and 2 blues both have three different paths leading to that total. So we have a 'multiplier' of $x3$:

\[
P(1 \text{ blue}) = P(X = 1) = \left(\frac{3}{4}\right)^2 \left(\frac{1}{4}\right) \times 3 \approx 0.1406 \\
P(2 \text{ blues}) = P(X = 2) = \left(\frac{3}{4}\right) \left(\frac{1}{4}\right)^2 \times 3 \approx 0.4219 \\
\]

Notice that $0.0156 + 0.1406 + 0.4219 + 0.4219 = 1$

**Function**

Above is just an example for 3 trials. We need to know how calculate probabilities for any $n$. The multiplier effect is illustrated in the $(r)$ part of the following function:

For $n$ trials, the probability that there are $r$ successes & $n-r$ failures is:

\[
P(X = r) = \binom{n}{r} p^r (1-p)^{n-r} \quad \text{for } r = 0, 1, \ldots, n
\]

E.g. = 8 rolls of a dice, prob. of rolling 5 sixes?

\[
\begin{align*}
\text{n} &= 8 \\
\text{r} &= 5 \\
p &= \frac{1}{6} \\
P(X = 5) &= \binom{8}{5} \left(\frac{1}{6}\right)^5 \left(\frac{5}{6}\right)^3 = 56 \times \left(\frac{1}{6}\right)^5 \times \left(\frac{5}{6}\right)^3 \approx 0.00417 \\
\end{align*}
\]

See 1.3
5.9 - NORMAL DISTRIBUTION

When you have a large single set of values, a histogram/bar chart becomes unfeasible. So an alternative is to approximate this into one bell-shaped curve.

**Properties**
- Symmetrical Bell-Shaped Curve
- The area under the curve is equal to 1 (or 100%)
- It is defined by its mean (μ) & standard dev. (σ)
- The mean is in the center
- ~ 68% of the data is within 1σ of the mean
- ~ 95% is ±2σ of the mean, ~ 99% is ±3σ.

**General Diagram**

**Normal Prob. Calculations**
When you are given the μ & σ of a curve, you can find the probability that a randomly selected value would lie within a certain range of x:

**Example**
- A set of 2000 IQ scores is normally distributed with μ = 100 & σ = 10. Find the probability of picking an IQ between 80 & 110.

**Solution** (TI-nspire)
- Enter L.B. = 80, U.B. = 110, μ = 100, σ = 10
- Answer will be shown as decimal
- Prob. = 0.819 = 81.9%

(TI-84)
- 2nd [DISTR][2: normal cdf]
- Rest is the same as above...

**Inverse Norm. Calculations**
This style of question is the opposite. You will be given an area/prob., and asked to find a boundary:

**Example**
The vol. of milk cartons has μ = 995 ml & σ = 5 ml. 10% of cartons are < x ml. Find x:

**Solution** (TI-nspire)
- Enter Area = 0.1, μ = 995, σ = 5
- The answer (998.6 ml) will be an upper boundary. To do find a lower boundary, do [1 - answer].

(TI-84)
- 2nd [DISTR][3: InvNorm][ENTER]
- Rest is the same as above...

**Z-Distribution**
In certain cases, it is helpful to work with a normal distribution with μ = 0, σ = 1. This is called the Z-distribution, or Z ~ N(0,1).

**Method**
To use the Z-dist., you must also transform each x-value to what we call a Z-score.

\[ Z = \frac{x - \mu}{\sigma} \]

(This also represents how many S.D.'s from μ it is)