GUIDE TO DATA PROCESSING

1. DETERMINING TYPE OF PROCESSING

In science, descriptive statistics are used to summarise key features of a dataset, to make information meaningful. This includes measures of central tendency (mean, median, mode) and spread (standard deviation, standard error, range). Different statistics may be used according to the data type employed and are presented via different types of graphs. Nominal data (categories) is presented via bar graphs, while numerical data (intervals) is presented via scatter plots. Mean and standard deviation are used when the data is normally distributed (bell-shaped distribution), while median and interquartile range are utilised when the spread of data is asymmetrical. The symmetry of a data spread is determined via a test of *skewness*. A skewness value between –2 and +2 is acceptable for parametric statistics (using the mean), whereas values outside this range suggest non-normality (use median **or** discuss validity).

Data can be processed using **Microsoft Excel**. Typically, the independent variable is placed in the first column, while each trial of the dependent variable is placed in subsequent columns. Formula for data processing are inputted by use of an equation symbol (=), followed by a command and then the rows to be processed (selected by highlighting). Once inputted, the formula can be applied to subsequent rows of data by clicking on the formula cell and then dragging the small box in the bottom right corner.

	A	B	C	D	E	F	G	H
1	1	13	16	12	12	15	14	= AVERAGE (B1:G1)

KEY: ■ INDEPENDENT VARIABLE ■ DEPENDENT VARIABLE DATA CLICK AND DRAG

Skewness:

	A	B	C	D	E	F
1	13	16	12	12	15	14
2	15	16	15	17	25	17

Formula: = SKEW (A1:F1)
0.383 → Acceptable skew
2.147 → Non-normality

Normal Spread:

	A	B	C	D	E	F
1	13	16	12	12	15	14

Mean:	**Std Dev:**
= AVERAGE (A1:F1)	= STDEV (A1:F1)

Nonparametric:

	A	B	C	D	E	F
2	15	16	15	17	25	17

Median:	**IQR:**
= MEDIAN (A2:F2)	= IQR (A2:F2)

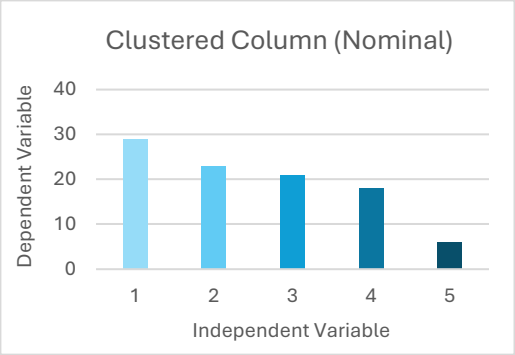
Helpful tips:

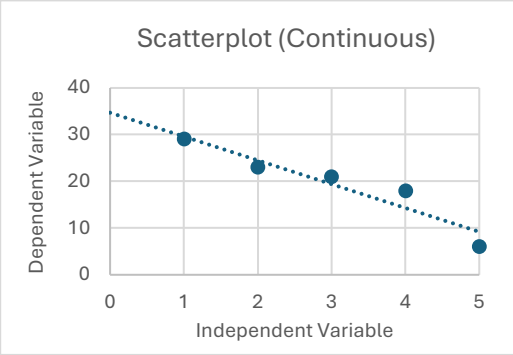
- Cells must only include numbers (no letters / spaces) – Excel cannot process non-numerical data
- All data should be processed the same way (either mean or median) – do not use interchangeably
- If a single data set is skewed, use mean and discuss the skewness as part of the analysis of data

2. GRAPHING

Processed data can be represented graphically by creating a new data set containing the independent variable and the averaged values for the dependent variable. To copy numbers generated by a formula, copy the relevant cells and then select PASTE > VALUES. To generate a graph, highlight the data set and select INSERT > CHART. Use a clustered column (nominal data) or an X-Y scatterplot (continuous data).

	A	B
1	1	29.8
2	2	25.2
3	3	21.4
4	4	17.5
5	5	13.7





Generating a graph will create a 'Chart Design' tab, enabling a user to add chart elements. This allows for horizontal and vertical axes titles to be created, along with the insertion of a trendline (scatterplot). Axes can be formatted to change the size of intervals, and the colour of columns can be customised. Different types of trendlines can be chosen to best suit the data spread (via 'More Trendline Options').

3. INTERPRETING STRENGTH OF RELATIONSHIPS

The strength of the relationship between the independent variable and the dependent variable can be represented using a **correlation coefficient** (r value). An r value shows the strength and direction of a *linear* relationship. A positive value indicates that both variables increase together, while a value that is negative indicates that the variables move in opposition. The size of the r value can vary between 1 (a perfect correlation) and 0 (no correlation). An r value of less than 0.5 is considered relatively weak.

If the relationship between the variables is not linear, a **coefficient of determination** (r^2 value) can be used to show how well the data fits the trendline selected. Unlike a correlation coefficient, the r^2 value does not indicate direction (it is always a positive value). The size of the r^2 value reflects the predictive power of the trendline (values range from 0 to 1). A value of 0.8 indicates that 80% of all changes in the dependent variable are likely related to changes in the independent variable, with 20% of changes due to other factors. It is important to recognise that neither coefficient will function to prove a causation.

	A	B	C	D	E
1	1	2	3	4	5
2	29	23	21	18	6

KEY: ■ INDEPENDENT ■ DEPENDENT (MEAN)

Correlation Coefficient:	Coefficient of Determination:
= CORREL (A1:E1 , A2:E2)	= RSQ (A1:E1 , A2:E2)

NOTE: The r^2 value can be displayed under trendline options

In the example shown above, the correlation coefficient (r value) is –0.95, indicating a strong negative correlation. The coefficient of determination (r^2 value) for a linear trendline is 0.90, indicating that 90% of the variation in the dependent variable is explained by the independent variable (good trendline fit).

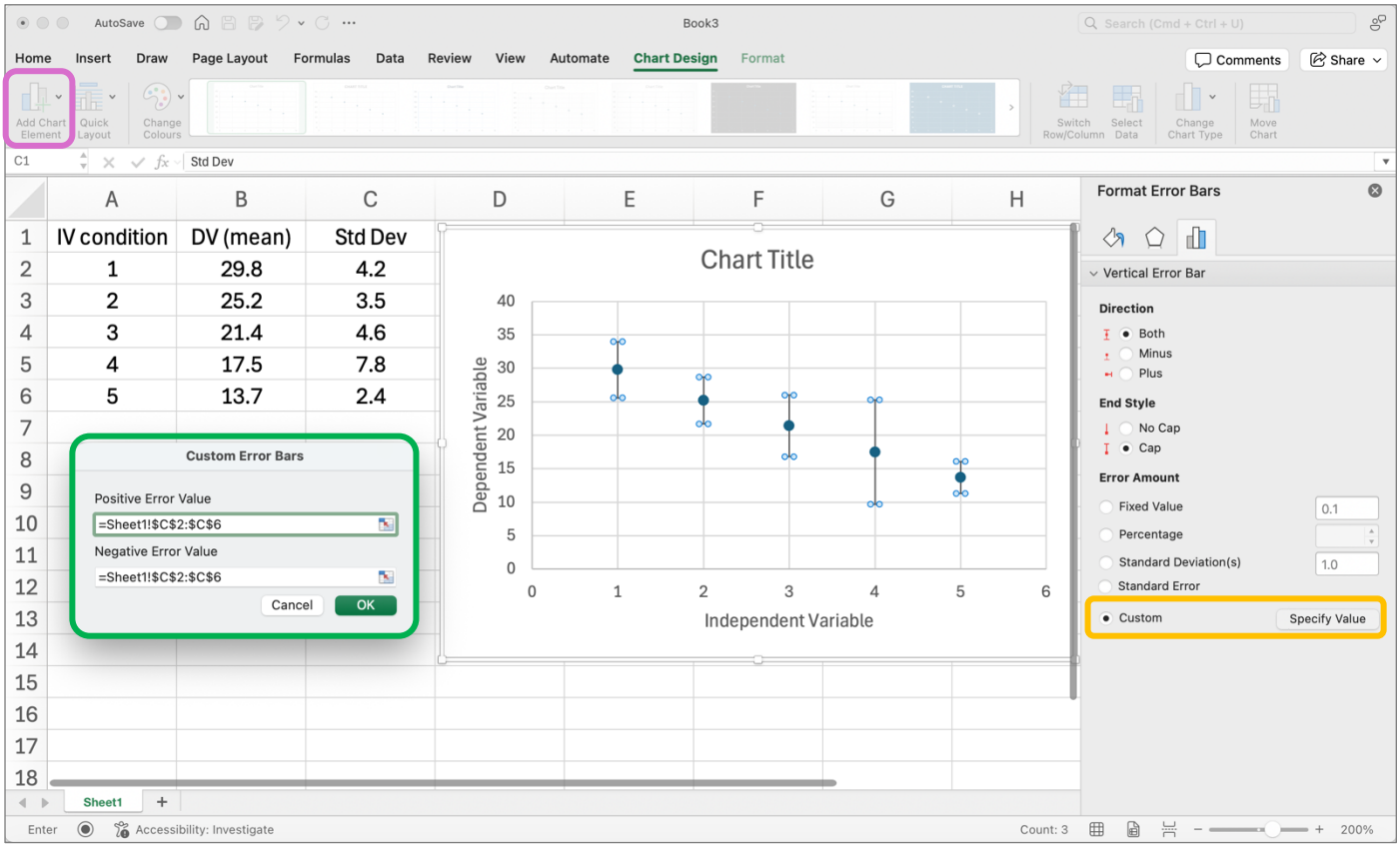
4. INTERPRETING PRECISION

Precision refers to the degree of similarity between multiple measurement values. It acts to provide an indication of the consistency and reliability of an experimental process. High precision means that there is very little variation between repeated data recordings. Precision is affected by *random errors*, which cause unpredictable variations in the measurement process. These errors cannot be avoided, and their displacement of data values is not consistent. The impact of random error can be limited by keeping all extraneous conditions constant (i.e. by maintaining control variables) to ensure precision.

Precision can be represented graphically via **error bars**. Standard deviation can be used as a measure of data spread in experiments with a minimum of five trials (if less than five, use the arithmetic range). When interpreting error bars, a key consideration is whether the error bars of two adjacent data points overlap (as this would suggest the difference *within* a data point is more significant than the difference *between* two data points). Overlapping error bars suggest the data trend between two points may not be significant (this can subsequently be confirmed by performing certain inferential statistical tests).

Error bars can be added to the graph by selecting CHART DESIGN > ADD CHART ELEMENT > ERROR BARS. To create error bars reflecting the standard deviation, a user must select the 'More Error Bars Option...' and then under 'Error Amount' select the custom option to specify specific values. Under the positive and negative error value, the cells containing the calculated standard deviation values should then be selected. This will result in the creation of variably sized error bars that represent standard deviation.

Including Error Bars in Excel:



- Select 'Add Chart Element' to access error bar options (under the 'Chart Design' tab)
- Select 'Custom' and then 'Specify Value' under the error amount (in 'Format Error Bars')
- Highlight the cells containing standard deviation for both positive and negative error values

5. INTERPRETING ACCURACY

Accuracy refers to the degree of similarity between a measurement value and a reference (or a 'true') value. Accuracy represents the exactness of a measurement, and the determination of accuracy will therefore require the prior establishment of the correct measurement standard. Any quantitative data should always be more accurate than qualitative data. Accuracy is impacted by *systematic errors* that cause predictable variations in a measurement process (displacement is consistent and directional). Sources of systematic errors include faulty device calibrations and faulty readings (e.g. parallax error).

The degree of inexactness in a measurement can be represented as an **uncertainty**. A manufacturer will commonly provide the level of absolute uncertainty for a device (± amount), otherwise a user may treat the uncertainty as being the minimum value for a digital device or half the minimum value for an analogue device. The impact of an uncertainty is proportionate to the size of the quantity that is being measured (a larger quantity will result in a smaller relative percentage). The overall level of inaccuracy for all measuring devices combined can be determined by a user via the total percentage uncertainty.

- $\boldsymbol{Total\ percentage\ uncertainty} = (Absolute\ uncertainty \div Quantity\ measured) \times 100$

Calculating Percentage Uncertainty:

	A	B	C	D	E	F	G
1	Absolute Uncertainty	0.5	0.1	1	0.25	0.4	**Total Percentage Uncertainty:**
2	Quantity Measured	3.4	3.0	60	2.75	8.0	
3	**% Formula:** = A1/A2*100	14.7	3.3	1.7	9.1	5	**= SUM (B3:F3)**

It is important to recognise that the absolute uncertainties of devices and equipment are not the only potential source of inaccuracy. A person measuring time would have a level of inaccuracy according to their reaction time – this cannot be quantitated but must be considered when assessing accuracy.

Equipment Uncertainties:

Volume:
- 250ml measuring cylinder: ±1ml
- 100ml measuring cylinder: ±0.6ml
- 50ml measuring cylinder: ±0.4ml
- 25ml measuring cylinder: ±0.3ml
- 10ml measuring cylinder: ±0.1ml

Fluid Dispensers:
- 1ml syringe: ±0.01ml
- 5ml syringe: ±0.1ml
- 3ml plastic pipette: ±0.2ml
- 200µl micropipette: ±0.5µl
- 1000µl micropipette: ±0.6µl

Temperature:
- Water bath: ±0.25ºC
- Incubator: ±0.24ºC
- Refrigerator: ±0.5ºC
- Thermometer: ±0.5ºC

Digital Devices:
- Pasco CO_2 sensor: ±100ppm*
- Pasco Colourimeter: ±0.001AU**
- pH meter: ±0.2pH
- Stopwatch: ±0.01sec
- Electronic balance: ±0.01g
- Digital Vernier Calliper: ±0.01mm

* Inaccurate below 1000ppm (Palmer, 2025)

** Inaccurate above 2.0 AU (Vernier Science, 2020)

6. DETERMINING STATISTICAL SIGNIFICANCE

Statistical tests are used to determine if observed patterns and trends in sample data are likely due to chance or are representative of a true effect. Different tests are employed according to the data type. Two common statistical tests for parametric data are the T-test (t-value) and the ANOVA test (f-value).

6.1 T–TEST

T-tests compare the means of two sets of data to determine if there is a significant difference between them. T-tests can be unpaired (two distinct groups) or paired (same group under different conditions). Two hypotheses are tested – the null hypothesis predicts no difference, while the alternate hypothesis predicts a difference between the two data sets. The null hypothesis is rejected if the t-value exceeds a critical threshold representing a p-value of 0.05 (indicating there is less than a 5% probability results are due to chance). A two-tailed test predicts a difference, while a one-tailed test predicts a difference *and* a direction. If a linear trendline is produced, a control group can be repeatedly compared to each experimental group to determine the point at which the trendline will become statistically significant.

A t-test can be performed using the DATA ANALYSIS option under the 'Data' tab. After selecting the right test (e.g. 't-Test: Two-Sample Assuming Equal Variances'), the two data sets need to be added into the input ranges. The target p-value can also be selected (default is 0.05). Excel will identify the generated t-value, the associated p-value and the critical t-value that must be exceeded to get a value of $p<0.05$. The critical value is dependent on the sample size of the two data sets (N_1 and N_2), as this determines the degree of freedom ($df = N_1 + N_2 - 2$). Each sample must have a minimum of five values to be valid.

Example: Two-Sample T-test Assuming Equal Variances

Control Group	Experimental
1	5
5	4
3	3
2	1
6	6
4	8

T-Test Data (p<0.05)	
Degree of freedom (df)	10
T-value (t Stat)	**0.7989**
P-value (one-tailed)	*0.221*
T-critical (one-tailed)	**1.812**
P-value (two-tailed)	*0.443*
T-critical (two-tailed)	**2.228**

The t-value does not exceed the critical threshold for either a one-tailed or two-tailed t-test. Therefore the p-value is greater than 0.05 and the null hypothesis is accepted (alternative hypothesis rejected).

Example: Critical values for t-test

Significance levels: One-tailed				
df	*0.1*	***0.05***	*0.025*	*0.01*
8	1.397	1.860	2.306	2.896
9	1.383	1.833	2.262	2.821
10	1.372	1.812	2.228	2.764

Significance levels: Two-tailed				
df	*0.1*	***0.05***	*0.025*	*0.01*
8	1.860	2.306	2.752	3.355
9	1.833	2.262	2.685	3.250
10	1.812	2.228	2.634	3.169

6.2 ANOVA (ANALYSIS OF VARIANCE)

ANOVA tests compare the means of multiple sets of data (i.e. more than two) to determine if there will be a significant difference between them. The F-ratio that is generated will indicate the impact that an independent variable has on the dependent variable (the larger the ratio, the greater the impact), but it does not determine the difference between specific sets of data. If the F-ratio exceeds a threshold for significance ($p<0.05$), a post-hoc test is conducted to pinpoint which specific groups show a relevant difference. If sample sizes are equal, a Tukey's HSD test is employed (honestly significant difference).

Microsoft Excel can perform an ANOVA test but lacks the native capacity to undertake an associated post-hoc test. However, a resource pack can be downloaded* to enable this function. Once installed, the tool can be activated by selecting 'Analysis Tools' in the Data tab and then enabled via a shortcut (**control-m**). For most experiments, an ANOVA: Single Factor test the follow-up option of a Tukey HSD is the most appropriate selection. When performing an ANOVA test, the raw data should be organised into columns, with each condition labelled for ease of comparison. The results should be presented in a table that is colour-coded, while pair-wise comparisons can be shown as letters on a graph – if two groups share a letter, then they are not significantly different from each other according to Tukey HSD. Letters are added to a graph by selecting the 'Data Labels' option under the 'Add Chart Element' tab.

Example: One-way ANOVA with post-hoc Tukey HSD

The example provided below uses data from an experiment in which six different concentrations of an antibiotic (0mg, 1mg, 2mg, 3mg, 4mg, 5mg) were tested to determine the impact on bacterial growth (as measured by the zone of inhibition). The data shows statistically significant differences between nearly all concentration pairs – except between 0mg to 1mg ($p = 0.740$) and 4mg to 5mg ($p = 0.604$).

ANOVA Results:

F-ratio: 112 ($\alpha = 0.05$)

F-critical: 2.53 ($df = 5,30$)

p-value: 1.11×10^{-16}

<u>To work out *df* values</u>:

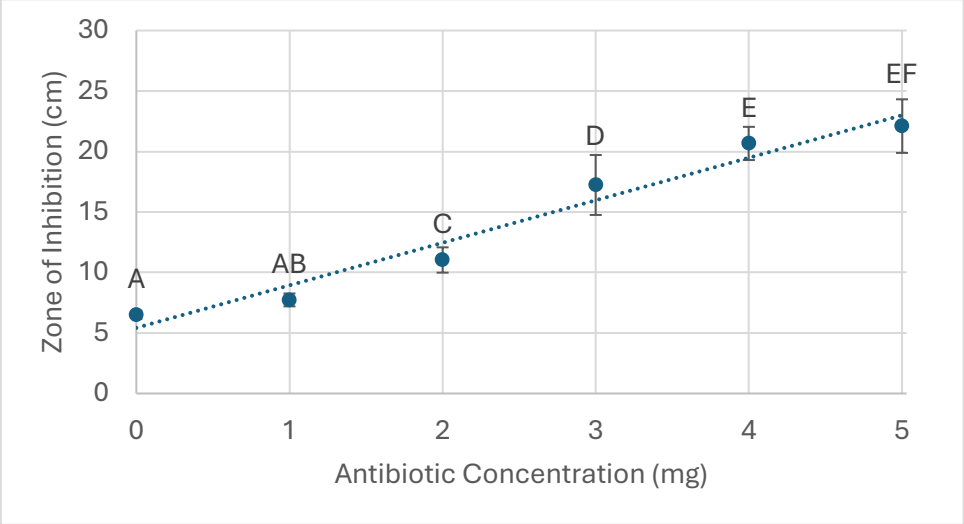
df (Between) = k – 1

df (Within) = n – k

Where:

n = number of data points

k = number of groups



Tukey HSD Results:

Group	1 mg	2 mg	3 mg	4 mg	5 mg
0 mg	0.734	2.59×10^{-4}	1.22×10^{-11}	3.86×10^{-12}	3.85×10^{-12}
1 mg	–	0.011	1.68×10^{-10}	3.93×10^{-12}	3.86×10^{-12}
2 mg	0.011	–	1.54×10^{-4}	1.22×10^{-10}	7.74×10^{-12}
3 mg	1.68×10^{-10}	1.54×10^{-4}	–	7.18×10^{-3}	9.23×10^{-5}
4 mg	3.86×10^{-12}	1.22×10^{-10}	7.18×10^{-3}	–	0.604

*** Analysis Tool pack:** https://real-statistics.com/free-download/real-statistics-resource-pack/