SCI1020 – Exam Summary

SCI1020 expands substantially on the statistics concepts above, particularly **data description, regression, sampling, inference, study design and statistical testing**.

1. Describing Data

For quantitative data, describe:

Centre

- Mean
- Median
- Mode

Spread

- Range = maximum − minimum
- Standard deviation
- IQR = Q3 − Q1

Shape

- Symmetrical/normal
- Positively skewed
- Negatively skewed
- Bimodal
- Outliers

Outlier boundaries:

$$Q1 - 1.5(IQR)$$
$$Q3 + 1.5(IQR)$$

For skewed data, **median + IQR** are generally more resistant to outliers than mean + SD.

2. Histograms, Boxplots & Normal Distribution

Histogram

- Shows distribution/frequency of quantitative data.

Boxplot
Five-number summary:

- Minimum
- Q1

- Median
- Q3
- Maximum

Use boxplots to compare:

- Centre
- Spread
- Skewness
- Outliers.

Normal distribution

- Symmetrical
- Mean = median = mode
- Spread determined by SD.

68–95–99.7 rule:

- ~68% within **1 SD**
- ~95% within **2 SD**
- ~99.7% within **3 SD**

3. Correlation & Regression

Regression equation:

$$\hat{y} = a + bx$$

- a= intercept
- b= slope.

Slope: expected change in Y for every 1-unit increase in X.

R^2: proportion of variation in **Y explained by the regression model**.

For a linear model, residuals should appear randomly scattered around zero.

Important:

- Don't extrapolate far outside the observed X-range.
- Correlation ≠ causation.
- Consider lurking/confounding variables.
- A curved residual pattern suggests a **non-linear relationship**.

4. Linear, Exponential & Power Relationships

Linear

- Constant additive change.

Exponential

- Constant multiplicative/proportional change.
- Can have constant doubling or half-life.

Exponential relationships can be linearised using:

$$\ln(y) \text{ vs } x$$

Power relationships:

$$y = Ax^b$$

can be linearised using:

$$\ln(y) \text{ vs } \ln(x)$$

5. Population, Sample, Parameter & Statistic

Population: entire group of interest.

Sample: subset of the population.

Parameter: numerical characteristic of the **population**.

Statistic: numerical characteristic calculated from a **sample**, used to estimate a population parameter.

Law of large numbers

As sample size increases:

sample estimate → tends to become closer to the true population value.

Standard error

Measures variability of a statistic between samples.

For the sample mean:

$$SE = \frac{s}{\sqrt{n}}$$