# Assignment Solutions – Unit 2 – Statistical Analysis

Model answers:

1. What is a box plot? List its key components.

A box plot is a graphical summary of a dataset that shows its distribution through five numbers: minimum, first quartile (Q1), median (Q2), third quartile (Q3) and maximum. The ‘box’ represents the interquartile range between Q1 and Q3, the line inside the box marks the median, and the ‘whiskers’ extend to the minimum and maximum values. Outliers may be plotted as individual points.

1. Define correlation and regression analysis. How are they different?

Correlation is a statistical measure that expresses the extent to which two variables linearly relate to one another; it is unitless and ranges from −1 to +1. Regression analysis, on the other hand, estimates the functional relationship between a dependent variable and one or more independent variables, often used to predict the dependent variable. While correlation quantifies association, regression provides an equation for prediction.

1. Explain the importance of data cleaning and imputation techniques.

Real‑world datasets often contain missing, duplicate or inconsistent values. Data cleaning identifies and fixes these issues to ensure analysis is accurate. Imputation is the process of estimating missing values using statistical methods such as mean substitution, median substitution or regression. Proper cleaning and imputation reduce bias and improve the validity of results.

1. Briefly describe ANOVA and the Chi‑Square test.

Analysis of Variance (ANOVA) tests whether there are statistically significant differences between the means of three or more groups by comparing variances within and between groups. The Chi‑Square test is used for categorical data to determine if there is a significant association between observed and expected frequencies. Both tests help evaluate hypotheses about relationships in data.

1. Explain what a scatter diagram is and how it aids in regression analysis.

A scatter diagram is a graph of paired data points (x,y) that shows how one variable relates to another. Each point represents one observation. Patterns in the plot may reveal linear or non‑linear relationships. In regression analysis, scatter plots help identify whether a linear model is appropriate and highlight outliers or clusters before fitting a regression line.

1. Describe the concept of hypothesis testing, including null and alternative hypotheses.

Hypothesis testing is a procedure in statistics to assess whether a belief about a population parameter is supported by sample data. The null hypothesis (H0) represents the default assumption (e.g. no difference or effect). The alternative hypothesis (H1) represents the claim researchers seek evidence for. Data are collected and a test statistic is calculated; if the statistic falls in a critical region determined by the significance level, H0 is rejected in favour of H1.