Probability and Statistics

Module 3: Descriptive Statistics

Complete classroom and examination notes

B.Sc. Computer Science • Course Code DASH250011

Based on the NIELIT Deemed to be University 2025 curriculum

Contents

1 Module map 1
2 1. Descriptive statistics and data types 1
2.1 Frequency distributions 1
3 2. Measures of central tendency 1
3.1 Arithmetic mean 1
3.2 Weighted mean 2
3.3 Median 2
3.4 Mode 2
3.5 Choosing a location measure 2
4 3. Measures of dispersion 2
4.1 Range 2
4.2 Interquartile range 3
4.3 Mean absolute deviation 3
4.4 Variance 3
4.5 Standard deviation 3
5 4. Moments 3
6 5. Skewness 4
7 6. Correlation 4
7.1 Covariance 4
7.2 Pearson correlation coefficient 5
7.3 Correlation cautions 5
8 7. Simple linear regression 5
8.1 Least-squares estimates 5
8.2 Meaning of the coefficients 6
8.3 Link between regression and correlation 6
8.4 Regression assumptions and diagnostics 6
9 8. Applications in business analytics and performance modelling 6
10 9. Problem-solving checklist 7
11 10. Quick formula sheet 7
12 11. Practice questions with answers 7

1 Module map

The syllabus module covers measures of central tendency—mean, median, and mode; measures of dispersion—range, variance, and standard deviation; moments and skewness; correlation and linear regression; and applications in business analytics and performance modelling.

> **Learning outcomes.** You should be able to summarize a dataset using appropriate numerical measures, distinguish location from spread, calculate and interpret moments and skewness, measure linear association using correlation, fit and interpret a simple linear regression model, and communicate the limitations of descriptive statistics.

2 1. Descriptive statistics and data types

Descriptive statistics organizes, summarizes, and presents observed data. It describes the sample at hand; it does not by itself prove that a relationship is causal or that a sample represents a wider population. A dataset may be:

- **Qualitative / categorical**: labels such as browser type, department, or defect class.
- **Quantitative / numerical**: measured or counted values.
- **Discrete**: countable values such as number of incidents.
- **Continuous**: measurements such as response time or CPU utilization.

A **population** is the complete group of interest. A **sample** is the observed subset. Population parameters are commonly written with Greek letters, while sample statistics use symbols such as $x \mid$ and s^2.

2.1 Frequency distributions

A frequency table records each value or class interval and its count f. The relative frequency is $\frac{f}{n}$, and the cumulative frequency is the running total of frequencies. For grouped data, the class midpoint is

$$m_i = \frac{\text{lower limit} + \text{upper limit}}{2}.$$

Approximate grouped-data mean:

$$x \mid \approx \frac{\sum_i f_i m_i}{n}.$$

For a histogram, use adjacent bars whose area represents frequency. A bar chart is generally used for categorical or discrete values; a histogram is used for continuous values grouped into intervals.

3 2. Measures of central tendency

Central tendency describes a typical or central location of a dataset.

3.1 Arithmetic mean

For observations $x_1, x_2, \ldots, x_n$, the sample mean is

$$x \mid = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}.$$

For a frequency distribution:

$$x \mid = \frac{\sum_i f_i x_i}{\sum_i f_i}.$$

The mean uses every observation and is useful for symmetric numerical data. It is sensitive to outliers.

> **Worked example: mean**
> For response times $12, 15, 15, 18, 20$ ms, $x\mid = \frac{12+15+15+18+20}{5} = 16$ ms.

3.2 Weighted mean

If observations have weights w_i, the weighted mean is

$$x\mid_w = \frac{\sum_i w_i x_i}{\sum_i w_i}.$$

Weights may represent importance, frequency, credit, or exposure. Always state whether the weights are normalized; normalization is not needed because the denominator rescales them.

3.3 Median

The **median** is the middle observation after sorting the data. If n is odd, it is the observation at position $\frac{n+1}{2}$. If n is even, it is the average of the observations at positions $\frac{n}{2}$ and $\frac{n}{2}+1$.
The median is resistant to extreme values and is often preferred for skewed data such as income, file size, or network latency.

3.4 Mode

The **mode** is the most frequently occurring value. A dataset can be unimodal, bimodal, multimodal, or have no unique mode. Mode is especially useful for categorical data, where mean and median may not be meaningful.

3.5 Choosing a location measure

Measure	Strength	Caution
Mean	Uses every numerical value; algebraically convenient.	Sensitive to outliers and skewness.
Median	Robust to outliers; good for skewed data.	Less convenient for algebraic calculations.
Mode	Works for categorical values; identifies most common case.	May not be unique or informative.

4 3. Measures of dispersion

Dispersion describes how far observations spread around a centre. Two datasets can have the same mean but very different variability.

4.1 Range

The range is

$$R = x_{\max} - x_{\min}.$$

It is simple and easy to interpret, but depends only on two observations and is highly sensitive to outliers.

4.2 Interquartile range

The first quartile Q_1 is the 25th percentile and the third quartile Q_3 is the 75th percentile. The interquartile range is

$$\mathrm{IQR} = Q_3 - Q_1.$$

The IQR describes the spread of the middle 50% and is resistant to extreme observations. A common boxplot outlier rule flags values below $Q_1 - 1.5\ \mathrm{IQR}$ or above $Q_3 + 1.5\ \mathrm{IQR}$.

4.3 Mean absolute deviation

About a centre c, the mean absolute deviation is

$$\mathrm{MAD} = \frac{1}{n}\sum_{i} |x_i - c|.$$

Often $c = x\ |$ or the median. Absolute deviations are easier to interpret than squared deviations but less convenient for algebra.

4.4 Variance

Population variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)^2.$$

Sample variance:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - x\ |)^2.$$

The $n-1$ denominator is Bessel's correction; it makes the sample variance an unbiased estimator of population variance under standard random-sampling assumptions.

The computational identity is

$$s^2 = \frac{\sum_i x_i^2 - n x\ |^2}{n-1}.$$

4.5 Standard deviation

The standard deviation is the positive square root of variance:

$$s = \sqrt{s^2}.$$

It has the same units as the original variable, which makes it more interpretable than variance. For approximately bell-shaped data, the empirical rule places about 68%, 95%, and 99.7% within 1, 2, and 3 standard deviations of the mean respectively.

> **Worked example: variance**
>
> For $2, 4, 4, 6$, $x\ | = 4$. Squared deviations are $4, 0, 0, 4$, whose sum is 8. The sample variance is $s^2 = \frac{8}{4-1} = \frac{8}{3} \approx 2.667$, and $s \approx 1.633$.

5 4. Moments

Moments provide a systematic description of the shape of a distribution.

The r-th raw moment about the origin is

$$\mu'_r = \frac{1}{n}\sum_i x_i^r.$$

The r-th central moment about the mean is

$$\mu_r = \frac{1}{n}\sum_i (x_i - x\ |)^r.$$

Important moments:

- First raw moment: $\mu'_1 = x\ |$.
- Second central moment: μ_2 is the population-style variance.
- Third central moment: related to asymmetry.
- Fourth central moment: related to tail heaviness and peakedness.

For sample work, use a clearly stated convention for denominators. Different textbooks may use corrected or uncorrected standardized moments.

6 5. Skewness

Skewness measures asymmetry around the centre.

- **Symmetric**: left and right sides have similar shape; mean and median are often close.
- **Right-skewed / positively skewed**: a long right tail; usually mean $>$ median.
- **Left-skewed / negatively skewed**: a long left tail; usually mean $<$ median.

A moment-based population coefficient is

$$\gamma_1 = \frac{\mu_3}{\sigma^3}.$$

A sample version commonly used is

$$g_1 = \frac{m_3}{s^3}, \quad m_3 = \frac{1}{n}\sum_{i(x_i - x\ |)}^{3}.$$

A Pearson-style descriptive coefficient is

$$\text{Skewness} \approx 3\frac{x\ | - \text{median}}{s}.$$

These formulas are not identical estimators. In an examination, use the formula specified by the question or textbook and state it.

> **Interpretation.** Skewness is about shape, not simply whether the mean is "large." A few unusually large values can pull the mean to the right and produce positive skewness.

7 6. Correlation

Correlation measures the strength and direction of linear association between two numerical variables X and Y.

7.1 Covariance

Sample covariance is

$$s_{XY} = \frac{1}{n-1}\sum_i (x_i - x\ |)(y_i - y\ |).$$

Positive covariance means that large values of X tend to occur with large values of Y. Negative covariance means that large values of X tend to occur with small values of Y. Its magnitude depends on measurement units, so covariance alone is difficult to compare.

7.2 Pearson correlation coefficient

The Pearson sample correlation is

$$r = \frac{\sum_i (x_i - x\ |)(y_i - y\ |)}{\sqrt{\sum^2_{i(x_i - x\ |)} \sum^2_{i(y_i - y\ |)}}}.$$

Equivalent form:

$$r = \frac{n\sum_i x_i y_i - \left(\sum_i x_i\right)\left(\sum_i y_i\right)}{\sqrt{\left(n\sum_i x_i^2 - \left(\sum_i x_i\right)^2\right)\left(n\sum_i y_i^2 - \left(\sum_i y_i\right)^2\right)}}.$$

Properties:

- $-1 \le r \le 1$.
- $r > 0$ indicates positive linear association; $r < 0$ indicates negative linear association.
- $|r|$ close to 1 indicates a strong linear pattern; $|r|$ close to 0 indicates weak linear association.
- Correlation is unit-free and unchanged by positive changes of scale.
- Correlation does not establish causation.

Worked interpretation

If $r = 0.86$ between CPU load and response time, there is a strong positive linear association in the observed data. It does not prove that CPU load alone causes response time, and it does not guarantee the same relationship outside the observed range.

7.3 Correlation cautions

- A curved relationship can have correlation near 0 even when variables are strongly related.
- A single high-leverage point can substantially change r.
- Restricting the range of one variable can weaken observed correlation.
- Confounding variables can create a spurious association.
- Always inspect a scatter plot before interpreting a correlation coefficient.

8 7. Simple linear regression

Simple linear regression models a response variable Y using one explanatory variable X:

$$Y = \beta_0 + \beta_1 X + \varepsilon.$$

Here β_0 is the intercept, β_1 is the slope, and ε is random error. In a fitted model, hats denote estimates:

$$\hat{y} = b_0 + b_1 x.$$

8.1 Least-squares estimates

The least-squares slope is

$$b_1 = \frac{\sum_{i(x_i - x\ |)} (y_i - y\ |)}{\sum^2_{i(x_i - x\ |)}}.$$

The intercept is

$$b_0 = y \mid - b_1 x \mid.$$

The fitted residual for observation i is

$$e_i = y_i - \hat{y}_i.$$

The least-squares line minimizes $\sum_i e_i^2$, the sum of squared residuals.

8.2 Meaning of the coefficients

- **Slope b_1**: expected change in predicted Y for a one-unit increase in X.
- **Intercept b_0**: predicted Y when $X = 0$; interpret only when $X = 0$ is meaningful and within the relevant range.
- **Residual**: observed value minus fitted value; positive means the model under-predicted.

> **Worked example: performance modelling**
>
> Suppose a fitted model predicts query time from database size: $\hat{y} = 12 + 0.8x$, where time is in ms and size is in GB. For $x = 10$, predicted time is 20 ms. The slope says each additional GB is associated with an increase of 0.8 ms in predicted time within the observed range.

8.3 Link between regression and correlation

For simple linear regression,

$$b_1 = r\frac{s_y}{s_x}.$$

The sign of the slope is the sign of r. When the regression includes an intercept, the coefficient of determination is

$$R^2 = r^2.$$

R^2 is the proportion of variation in Y explained by the fitted linear relationship with X in the sample. For example, $R^2 = 0.64$ means that 64% of the sample variation in Y is accounted for by the model's linear fitted values; it is not a causal percentage.

8.4 Regression assumptions and diagnostics

A basic linear-regression analysis should consider:

1. **Linearity**: the mean response changes linearly with X.
2. **Independence**: observations or errors are not dependent in an unmodelled way.
3. **Constant variance**: the spread of errors is approximately stable across fitted values.
4. **Normality of errors**: mainly important for small-sample confidence intervals and tests.
5. **No influential anomalies**: outliers and high-leverage points are investigated.

Use residual-versus-fitted plots to look for curvature or changing spread. Do not extrapolate far beyond the observed X values.

9 8. Applications in business analytics and performance modelling

Descriptive statistics supports fast, transparent decisions:

- **Service monitoring**: median latency, 95th percentile latency, standard deviation, and outlier rate.
- **Capacity planning**: mean and variability of requests per minute, CPU, memory, and queue length.
- **Business dashboards**: revenue mean, median transaction value, category mode, and quartiles.

- **Quality control**: defect rate, process spread, skewness, and correlation between process variables.
- **A/B analysis**: compare group means or medians while reporting sample sizes and dispersion.
- **Predictive baselines**: use simple regression to estimate performance from workload, data volume, or resource allocation.

A good report includes the data period, number of observations, units, missing-value handling, outlier policy, selected summaries, and an explanation of what the measures do and do not support.

10 9. Problem-solving checklist

1. Identify the variable types, units, sample size, and possible missing values.
2. Sort or tabulate data before finding median, quartiles, or mode.
3. State whether variance uses denominator n or $n-1$.
4. Report mean together with standard deviation when data are reasonably symmetric.
5. Report median together with IQR when data are skewed or contain outliers.
6. Draw or inspect a histogram, box plot, and scatter plot where appropriate.
7. For correlation, check linearity and influential points.
8. For regression, write the fitted equation and interpret slope in context.
9. Never turn correlation or a fitted line into a causal claim without a suitable study design.

11 10. Quick formula sheet

Concept	Formula
Arithmetic mean	$x \mid = \left(\frac{1}{n}\right) \sum_i x_i$
Range	$R = x_{\max} - x_{\min}$
Interquartile range	$\mathrm{IQR} = Q_3 - Q_1$
Sample variance	$s^2 = \left(\frac{1}{n-1}\right) \sum^2_{i(x_i - x \mid)}$
Standard deviation	$s = \sqrt{s^2}$
Third standardized moment	$\gamma_1 = \frac{\mu_3}{\sigma^3}$
Pearson correlation	$r = \sum_{i(x_i - x \mid)} \frac{y_i - y \mid}{\sqrt{\sum^2_{i(x_i - x \mid)} \sum^2_{i(y_i - y \mid)}}}$
Regression slope	$b_1 = \sum_{i(x_i - x \mid)} \frac{y_i - y \mid}{\sum^2_{i(x_i - x \mid)}}$
Regression intercept	$b_0 = y \mid - b_1 x \mid$
Coefficient of determination	$R^2 = r^2$ for simple regression with intercept

12 11. Practice questions with answers

1. For $4, 7, 7, 9, 13$, find the mean, median, mode, and range. **Answer:** mean $= 8$, median $= 7$, mode $= 7$, range $= 9$.
2. Why is the median preferred to the mean for a highly right-skewed income dataset? **Answer:** The median is less affected by a small number of very large incomes.
3. For $2, 4, 4, 6$, calculate the sample variance and standard deviation. **Answer:** $s^2 = \frac{8}{3} \approx 2.667$ and $s \approx 1.633$.

4. If $r = -0.72$ between cache size and response time, interpret the sign. **Answer:** Larger cache size is associated with lower response time in a moderately strong negative linear pattern; causation is not established.
5. A regression equation is $\hat{y} = 5 + 2.5x$. What does the slope mean? **Answer:** A one-unit increase in x is associated with a 2.5-unit increase in predicted y.
6. If $r = 0.8$ in a simple regression with an intercept, find R^2. **Answer:** $R^2 = 0.64$, or 64% of sample variation explained by the fitted linear relationship.
7. Name two reasons why a correlation near zero can be misleading. **Answer:** The relationship may be curved, or outliers/confounding/range restriction may hide the pattern.

> **Exam reminder.** Show the formula, substitution, and interpretation. A descriptive statistic is not complete until its unit and context are stated—for example, "median latency = 42 ms," not merely "median = 42."