DegreeLive

Bachelor of Commerce (B2001) | Level 1 • Part A Core | Monash University

Business and economic statistics

Course Code: ETC1000 • Module 1 Notes

Module 1: Data Summarisation, Graphical Methods, Central Tendency, Dispersion and Standardisation

1.1 Role of Statistics in Australian Business and Economic Decision-Making

Modern commerce, public policy, and financial management rely on systematic quantitative evidence rather than intuition alone. In Australian organisations—ranging from ASX-listed corporations and major retail banks regulated by the Australian Prudential Regulation Authority (APRA) to state treasury departments and small-to-medium enterprises (SMEs)—decision-makers routinely confront uncertainty in consumer demand, labour costs, interest rates, and supply-chain lead times. Business and economic statistics provides the formal methodology for collecting, organising, summarising, analysing, and interpreting numerical and categorical data so that managers and economists can convert raw operational records into defensible decisions.

Statistical inquiry is conventionally divided into two complementary branches. **Descriptive statistics** encompasses numerical, tabular, and graphical procedures that summarise and reveal patterns within an observed dataset without making formal claims beyond the observations at hand. When an operations analyst at a Melbourne supermarket chain computes average weekly basket spend, constructs a histogram of checkout wait times, or calculates the interquartile range of store-level electricity expenditure, the analyst is performing descriptive statistics. **Inferential statistics**, developed in subsequent modules, extends those summaries by using probability theory to draw conclusions, construct interval estimates, and test formal hypotheses about an unobserved broader group based on a subset of observations.

Populations, Samples, Parameters, and Statistics

Rigorous statistical reporting requires precise differentiation between the complete target group and the observed subset:

- **Population (N):** The entire collection of individuals, firms, transactions, or economic units about which information is sought. Examples include all residential mortgages written by Australian authorised deposit-taking institutions in a financial year, all actively trading businesses registered with an Australian Business Number (ABN), or every daily closing price of the S&P/ASX 200 index since inception.
- **Sample (n):** A finite subset of units selected from the population for measurement and analysis. Conducting a complete enumeration (a **census**, such as the five-yearly Australian Bureau of Statistics Census of Population and Housing) is often prohibitively expensive, slow, or destructive to the tested item, making probability sampling the standard commercial practice.
- **Parameter:** A fixed, typically unknown numerical summary characteristic of a population, conventionally denoted by Greek letters such as the population mean (μ), population variance (σ^2), population standard deviation (σ), or population proportion (p or π).
- **Statistic:** A numerical summary computed solely from sample observations, conventionally denoted by Latin letters such as the sample mean ($\bar{x}$), sample variance (s^2), sample standard deviation (s), or sample proportion ($\hat{p}$). Because a statistic depends on which random subset happens to be drawn, it varies across repeated samples.

Data Structures and Governance in Australia

Business and economic datasets arise in three primary structural dimensions. **Cross-sectional data** capture measurements across multiple entities at a single point or period in time, such as the quarterly revenue, employee headcount, and debt-to-equity ratio of 150 ASX-listed industrial firms at 30 June 2026. **Time-series data** track a single entity or aggregate economic indicator sequentially across equally spaced intervals, such as the Reserve Bank of Australia (RBA) monthly Cash Rate Target, quarterly Australian Consumer Price Index (CPI) inflation, or daily AUD/USD exchange rates. **Panel (longitudinal) data** combine both dimensions by following the same cross-sectional units across multiple time periods, such as tracking household disposable income and housing stress for 8,000 Australian households across ten annual waves of the Household, Income and Labour Dynamics in Australia (HILDA) survey.

Data collection is further classified as **observational** or **experimental**. Observational data arise from passive recording of market outcomes, administrative filings (such as Australian Taxation Office Business Activity Statements), or point-of-sale scanner logs without researcher intervention. Experimental data arise when an analyst deliberately manipulates a treatment variable under controlled conditions—for instance, an Australian e-commerce retailer randomly assigning website visitors to

DegreeLive

two checkout page layouts (A/B testing) to measure conversion rates. Regardless of source, commercial analysts must ensure sampling frames exclude systematic coverage bias and handle customer-level records in compliance with the Australian Privacy Principles under the *Privacy Act 1988 (Cth)*, while listed entities reporting quantitative financial metrics remain bound by continuous disclosure obligations under the *Corporations Act 2001 (Cth)*.

1.2 Classification of Data and Measurement Scales

Selecting an appropriate summary measure, chart, or statistical model depends directly on the nature of the recorded variable. Misapplying arithmetic operations—such as averaging numeric postcodes or treating ordinal survey ratings as strictly ratio-scaled—produces meaningless conclusions. Variables are classified along two orthogonal dimensions: data type (categorical versus numerical) and Stevens' four scales of measurement (nominal, ordinal, interval, and ratio).

Categorical versus Numerical Variables

Categorical (qualitative) variables assign each observation to one of several mutually exclusive categories or attributes. Examples include the state or territory of business registration (VIC, NSW, QLD, WA, SA, TAS, ACT, NT), Global Industry Classification Standard (GICS) sector on the ASX, or whether a commercial loan is in arrears (Yes/No). Even when categories are encoded numerically for database storage (e.g., 1 = Retail, 2 = Manufacturing, 3 = Mining), the digits act solely as labels.

Numerical (quantitative) variables represent genuine counts or physical/monetary measurements where arithmetic differences and averages are physically meaningful. Numerical variables split into two subtypes:

- **Discrete numerical variables** arise from counting processes and take isolated, countable values (typically non-negative integers) with gaps between adjacent possible values. Examples include the number of insurance claims lodged per day, the count of defective pallets in a Brisbane warehouse shipment, or the number of branches operated by a regional bank.
- **Continuous numerical variables** arise from measuring on an unbroken continuum and can theoretically assume any real value within an interval, constrained only by instrument precision. Examples include shipping transit time in hours, daily ASX turnover in A$ millions, lithium ore grade in percentage weight, or household electricity consumption in kilowatt-hours.

Stevens' Four Levels of Measurement

The four measurement scales form a strict hierarchy of mathematical information content, from lowest to highest:

Scale	Distinguishing Property	Australian Business Examples	Permissible Operations & Summaries
Nominal	Categories serve purely as labels with no intrinsic ranking or order.	ASX sector classification; payment method (EFTPOS, PayID, BPAY, Cash); Australian state of incorporation.	Equality ($=$ / $\neq$), frequency counts, proportions, mode, bar and Pareto charts.
Ordinal	Categories possess a meaningful rank order, but intervals between adjacent ranks are not necessarily equal.	S&P credit ratings (AAA, AA+, AA, A+); 5-point Net Promoter or customer satisfaction scores; risk tiers (Low, Medium, High).	Order relations ($<$, $>$), median, percentiles, interquartile range (on coded ranks), bar charts.
Interval	Constant, meaningful unit differences exist between values, but the zero point is arbitrary (not a true absence of the attribute).	Temperature in degrees Celsius ($0°C$ does not mean zero thermal energy); calendar years; standardised z-scores; IQ/aptitude index scores.	Addition and subtraction ($+$, $-$), arithmetic mean, standard deviation, variance, histograms, boxplots. Ratios (e.g., "$20°C$ is twice as warm as $10°C$") are invalid.
Ratio	Possesses equal units AND a true, non-arbitrary absolute zero representing complete absence of the quantity.	Annual revenue in A$; household income; unemployment rate (%); production output in tonnes; inventory holding days.	All arithmetic operations ($+$, $-$, $\times$, $\div$), geometric mean, coefficient of variation, meaningful ratios ("revenue doubled").

DegreeLive

1.3 Tabular Summarisation and Frequency Distributions

Raw operational logs containing hundreds or thousands of individual numbers cannot be interpreted directly by management. Tabular summarisation condenses raw records into structured distributions that reveal concentration, spread, and anomalies.

Categorical Frequency and Contingency Tables

For a categorical variable with k distinct classes, a **frequency distribution** lists each category alongside its absolute frequency (f_j, the count of observations in category j), its **relative frequency** (f_j/n), and its **percentage frequency** ($(f_j/n) \times 100\%$). Clearly, $\sum_{j=1}^{k} f_j = n$ and $\sum_{j=1}^{k} (f_j/n) = 1$.

When analysts examine two categorical variables simultaneously—such as borrower employment status (Full-time, Part-time, Self-employed) against loan repayment status (Current, Arrears)—they construct a two-way **contingency table (cross-tabulation)**. The interior cells display **joint frequencies** (or joint relative frequencies), the row and column totals yield **marginal distributions**, and dividing each cell by its respective row or column total produces **conditional distributions**. Careful inspection of conditional distributions across subgroups prevents **Simpson's paradox**, a statistical phenomenon in which a trend or performance advantage that appears within every disaggregated subgroup reverses or disappears when the subgroups are combined due to unequal group sizes (confounding).

Grouped Frequency Distributions for Numerical Data

For continuous or wide-ranging discrete variables, individual values rarely repeat often enough for an ungrouped table to be informative. Instead, observations are partitioned into k contiguous, mutually exclusive, and collectively exhaustive **class intervals**. Constructing an effective grouped distribution involves four disciplined steps:

1. **Select the number of classes (k):** Typically between 5 and 15 classes depending on sample size n. Too few classes obscure distributional shape; too many create erratic, sparse gaps. **Sturges' guideline** provides a principled starting point:

   $$k \approx 1 + 3.322 \log_{10}(n)$$

   Where: k is the suggested number of class intervals (rounded to a convenient integer), and n is the total number of observations in the dataset.

2. **Determine the class width (w):** Divide the data range by k and round up to a clean, readable number (such as multiples of 5, 10, 50, or 100):

   $$w \geq \frac{\max(x_i) - \min(x_i)}{k}$$

   Where: w is the uniform width of each class interval, $\max(x_i)$ is the largest observed value, $\min(x_i)$ is the smallest observed value, and k is the chosen number of classes.

3. **Establish unambiguous class boundaries:** Use continuous boundary notation such as $[L_j, U_j)$, meaning $L_j \leq x < U_j$, so borderline observations fall into exactly one interval without ambiguity.
4. **Tabulate frequencies and cumulative summaries:** Record the frequency f_j, relative frequency f_j/n, cumulative frequency ($CF_j = \sum_{i=1}^{j} f_i$), and cumulative percentage ($CF_j/n \times 100\%$).

Worked Numerical Example 1: Grouped Frequency Distribution for Melbourne Retail Sales

A retail analytics team records weekly gross sales (in A$ thousands) across $n = 40$ Melbourne suburban convenience outlets during a non-promotional trading week. The minimum store revenue is A$21.4 thousand and the maximum is A$78.9 thousand.

DegreeLive

Step 1: Determine k and w. Applying Sturges' guideline gives $k \approx 1 + 3.322 \log_{10}(40) = 1 + 3.322(1.6021) = 6.32$. Selecting $k = 6$ classes yields a minimum required width of:

$$w \geq \frac{78.9 - 21.4}{6} = \frac{57.5}{6} = 9.58$$

Where: 78.9 is the maximum weekly sales (A$ thousands), 21.4 is the minimum weekly sales (A$ thousands), and 6 is the target number of classes. Rounding up to $w = 10$ and starting at a clean lower boundary of 20 produces six equal intervals spanning $[20, 30)$ through $[70, 80]$.

Step 2: Tabulate frequencies and compute the grouped mean approximation. Let $m_j = (L_j + U_j)/2$ denote the class midpoint.

Weekly Sales (A$ '000s)	Midpoint (m_j)	Frequency (f_j)	Relative Freq. (f_j/n)	Cum. Freq. (CF_j)	Cum. Percentage	$f_j \cdot m_j$
$[20, 30)$	25	4	0.100	4	10.0%	100
$[30, 40)$	35	7	0.175	11	27.5%	245
$[40, 50)$	45	12	0.300	23	57.5%	540
$[50, 60)$	55	9	0.225	32	80.0%	495
$[60, 70)$	65	5	0.125	37	92.5%	325
$[70, 80]$	75	3	0.075	40	100.0%	225
Total	—	**40**	**1.000**	—	—	**1,930**

Step 3: Compute the grouped mean and interpret. When only grouped data are published (common in ABS demographic tables), the arithmetic mean is approximated by assigning every observation in class j its midpoint m_j:

$$\bar{x}_{\text{grouped}} = \frac{\sum_{j=1}^{k} f_j m_j}{n} = \frac{1,930}{40} = 48.25$$

Where: f_j is the frequency of class j, m_j is the midpoint of class j, $k = 6$ is the number of classes, and $n = 40$ is the total sample size. Thus, estimated average weekly revenue across the 40 outlets is A$48,250. Moreover, 57.5% of outlets generate less than A$50,000 per week, and the modal class (highest frequency) is $[40, 50)$ with 30.0% of all stores.

1.4 Graphical Methods and Visual Integrity

Well-constructed charts expose distributional shape, skewness, multimodality, temporal shifts, and bivariate associations far faster than raw tables. Each variable type requires a specific graphical form.

Charts for Categorical and Numerical Data

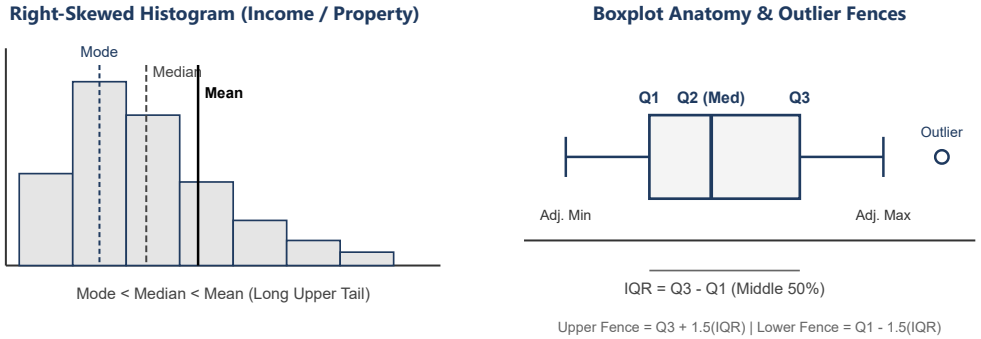
- **Bar Charts and Pareto Charts:** Display categorical frequencies or relative frequencies using separated rectangular bars of equal width. A **Pareto chart** arranges bars in descending order of frequency and overlays a cumulative percentage curve, allowing operations managers to isolate the "vital few" causes driving the majority of customer complaints or manufacturing defects.
- **Histograms:** Display grouped numerical distributions using adjacent rectangular bars with no gaps, reflecting the continuity of the underlying scale. When class widths are equal, bar height is proportional to frequency or relative frequency. When class widths are unequal—as frequently occurs in ATO income brackets—bar height must represent **frequency density** (f_j/w_j) so that bar *area* remains strictly proportional to frequency.

DegreeLive

- **Frequency Polygons and Ogives:** A frequency polygon connects class midpoints at their corresponding frequencies, facilitating overlay comparisons of two groups (e.g., Sydney versus Melbourne store sales). An **ogive** plots cumulative frequencies or cumulative percentages against upper class boundaries, allowing direct visual read-off of medians and percentiles.
- **Boxplots (Box-and-Whisker Plots):** Visualise the five-number summary (minimum, first quartile Q_1, median Q_2, third quartile Q_3, and maximum) alongside potential outliers. Boxplots are especially effective for comparing spread and skewness across multiple groups side by side.
- **Time-Series Line Plots and Scatter Plots:** Line plots connect observations ordered chronologically to reveal trends, seasonal cycles, and structural breaks. Scatter plots display paired numerical observations (x_i, y_i) to reveal linear or nonlinear relationships, clusters, and bivariate outliers.

Distributional Shape and Boxplot Anatomy

Inspecting histograms and boxplots reveals whether a distribution is symmetric, positively (right) skewed, or negatively (left) skewed. Economic variables such as Australian household income, executive remuneration, and residential property prices are almost universally **right-skewed**: a long upper tail of high values pulls the mean above the median. Conversely, exam score distributions with a high ceiling or asset-lifetime distributions subject to early wear-out may exhibit **left-skewness**, pulling the mean below the median.



Principles of Graphical Integrity

In corporate investor presentations and regulatory submissions, visual distortion can mislead stakeholders—conduct that in Australian commercial advertising may contravene Section 18 of the *Australian Consumer Law* (Schedule 2 of the *Competition and Consumer Act 2010 (Cth)*). Analysts must enforce three integrity rules: (1) bar charts must always begin their frequency or magnitude axis at zero, because truncating the baseline exaggerates small relative differences; (2) two-dimensional area or three-dimensional perspective effects must not be used for one-dimensional quantities, as doubling bar height and width quadruples visual area (violating Edward Tufte's lie-factor unity principle); and (3) time-series charts spanning multi-year inflationary periods should plot real (CPI-adjusted) rather than nominal dollar figures when comparing purchasing power.

1.5 Measures of Central Tendency (Location)

A measure of central tendency identifies a single representative value around which the observations cluster. Four measures dominate business and economic analysis: the arithmetic mean, the weighted mean, the geometric mean, and the median, complemented by the mode for categorical or discrete concentration.

The Arithmetic Mean and Its Properties

The **population arithmetic mean** (μ) and **sample arithmetic mean** ($\bar{x}$) sum all observations and divide by the number of observations:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad \text{and} \qquad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Where: μ is the population mean, N is the population size, $\bar{x}$ is the sample mean, n is the sample size, and x_i denotes the i-th numerical observation.

The arithmetic mean serves as the "centre of gravity" of the dataset because signed deviations around the mean always sum exactly to zero:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Where: $x_i - \bar{x}$ is the signed deviation of the i-th observation from the sample mean $\bar{x}$. Furthermore, under any linear transformation $y_i = a + bx_i$ (such as converting USD revenues to AUD via exchange rate b or adding a fixed administrative fee a), the sample mean transforms identically: $\bar{y} = a + b\bar{x}$. However, because every observation enters with equal weight $1/n$, the arithmetic mean is sensitive to extreme outliers.

The Weighted Mean

When observations carry unequal economic importance—such as asset classes in an Australian superannuation portfolio, stock prices in a market-capitalisation index, or expenditure items in the ABS Consumer Price Index basket—the **weighted mean** ($\bar{x}_w$) assigns each value x_i a non-negative weight w_i:

$$\bar{x}_w = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

Where: $\bar{x}_w$ is the weighted mean, x_i is the i-th value, and w_i is the weight assigned to the i-th value (when weights are portfolio proportions summing to 1, the denominator equals 1).

The Geometric Mean (Compound Growth Rate)

When averaging multi-period percentage returns, GDP growth rates, or price-index relatives that compound multiplicatively over time, the arithmetic mean systematically overstates actual accumulated wealth. The **geometric mean** (G) of n positive growth factors $(1 + r_t)$ yields the exact constant per-period compound growth rate (CAGR):

$$G = \left[\prod_{t=1}^{n} (1 + r_t)\right]^{1/n} - 1 = \sqrt[n]{(1 + r_1)(1 + r_2)\cdots(1 + r_n)} - 1$$

Where: G is the geometric mean rate of return per period, r_t is the decimal rate of return in period t (e.g., +0.082 for +8.2%), and n is the number of compounding periods.

The Median and the Mode

To obtain the **median**, observations are first sorted in ascending order $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$. The median lies at position $(n + 1)/2$:

DegreeLive

$$\text{Median} = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ is odd} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{if } n \text{ is even} \end{cases}$$

Where: $x_{(k)}$ denotes the k-th order statistic (the k-th smallest value in the sorted sample), and n is the sample size. Because the median depends only on middle rank positions and not on extreme magnitudes, it is **robust (resistant) to outliers**. Consequently, CoreLogic and the ABS report *median* Australian dwelling prices and *median* household income rather than arithmetic means.

The **mode** is the value or category occurring with the highest frequency. A dataset may be unimodal, bimodal (indicating two distinct underlying market segments, such as weekday commuter versus weekend leisure airfares), or possess no mode when all observations are unique. Unlike the mean and median, the mode is applicable to nominal categorical data.

Worked Numerical Example 2: Superannuation Returns and Portfolio Weighting

Part A (Arithmetic Mean, Median, and Geometric Mean): An Australian balanced superannuation option records annual net returns over five consecutive financial years of +8.2%, -3.5%, +11.4%, +6.8%, and +12.1%.

1. **Arithmetic Mean:**

$$\bar{r} = \frac{8.2 + (-3.5) + 11.4 + 6.8 + 12.1}{5} = \frac{35.0}{5} = 7.00\%$$

   Where: 8.2, -3.5, 11.4, 6.8, 12.1 are the five annual percentage returns and $n = 5$.

2. **Median:** Ordering the five returns gives $-3.5\%,\ 6.8\%,\ 8.2\%,\ 11.4\%,\ 12.1\%$. Since $n = 5$ is odd, the median sits at position $(5 + 1)/2 = 3$, giving 8.20%.

3. **Geometric Mean (CAGR):** Converting percentages to growth factors $(1 + r_t)$ yields 1.082, 0.965, 1.114, 1.068, 1.121. Multiplying these five factors gives $1.082 \times 0.965 \times 1.114 \times 1.068 \times 1.121 = 1.392569$ (a cumulative 5-year wealth growth of 39.26%). Taking the fifth root:

$$G = (1.392569)^{1/5} - 1 = 1.068472 - 1 = 0.068472 \approx 6.85\%$$

   Where: 1.392569 is the terminal cumulative wealth relative per A$1 invested, and $1/5$ reflects the five annual compounding periods. Notice that the geometric mean (6.85%) is strictly lower than the arithmetic mean (7.00%) due to return volatility (the -3.5% drawdown year).

Part B (Weighted Mean Portfolio Return): Suppose an investor allocates A$200,000 across four asset classes: 45% in Australian Equities (return 9.4%), 30% in Fixed Income (4.2%), 15% in A-REITs (6.8%), and 10% in Cash (4.1%):

$$\bar{x}_w = 0.45(9.4\%) + 0.30(4.2\%) + 0.15(6.8\%) + 0.10(4.1\%) = 4.23\% + 1.26\% + 1.02\% + 0.41\% = 6.92\%$$

Where: 0.45, 0.30, 0.15, 0.10 are the portfolio weights summing to 1.00, and 9.4%, 4.2%, 6.8%, 4.1% are the respective asset-class returns.

1.6 Measures of Dispersion, Position, and Shape

Central tendency alone cannot distinguish between a stable cash-flow stream and a highly volatile one. Two investment funds can share an identical 7.0% mean return while exhibiting vastly different downside risk. Measures of dispersion quantify variability around the centre.

DegreeLive

Range and Interquartile Range (IQR)

The **range** ($\max(x_i) - \min(x_i)$) measures total span using only the two most extreme values, making it unstable in the presence of outliers. A far more reliable non-parametric spread metric is the **interquartile range (IQR)**, defined as the span of the middle 50% of ordered observations:

$$\text{IQR} = Q_3 - Q_1$$

Where: Q_1 is the first quartile (25th percentile) and Q_3 is the third quartile (75th percentile). Quartiles are standardly computed from the ordered array $x_{(1)} \leq \cdots \leq x_{(n)}$ via linear interpolation at index $p(n+1)$ (or by splitting the ordered dataset at the median into lower and upper halves and taking the median of each half, Tukey's hinges—note that minor software conventions differ slightly for small n, and we state the exact convention used in each problem). Using the IQR, Tukey's **outlier detection fences** flag mild outliers outside:

$$\text{Lower Fence} = Q_1 - 1.5 \times \text{IQR} \qquad \text{and} \qquad \text{Upper Fence} = Q_3 + 1.5 \times \text{IQR}$$

Where: Q_1 is the 25th percentile, Q_3 is the 75th percentile, and $\text{IQR} = Q_3 - Q_1$.

Variance, Standard Deviation, and Bessel's Correction

The **population variance** (σ^2) averages the squared deviations of all population values from μ, while the **sample variance** (s^2) divides the sum of squared deviations from $\bar{x}$ by $n-1$:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad \text{and} \qquad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \frac{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}{n-1}$$

Where: σ^2 is the population variance, N is the population size, μ is the population mean, s^2 is the sample variance, n is the sample size, and $\bar{x}$ is the sample mean.

Using $n-1$ instead of n in the denominator of s^2 is known as **Bessel's correction**. Because $\bar{x}$ is estimated from the same sample and minimises the sum of squared deviations, $\sum(x_i - \bar{x})^2$ is systematically slightly smaller than $\sum(x_i - \mu)^2$. Dividing by $n-1$ (the **degrees of freedom**, reflecting the constraint $\sum(x_i - \bar{x}) = 0$ which leaves $n-1$ freely varying deviations) renders s^2 an *unbiased estimator* of σ^2, meaning $E(s^2) = \sigma^2$.

Because variance is measured in squared units (e.g., AUD^2), analysts take the positive square root to restore the original measurement units, yielding the **population standard deviation** ($\sigma = \sqrt{\sigma^2}$) and **sample standard deviation** ($s = \sqrt{s^2}$). Under any linear transformation $y_i = a + bx_i$, adding a constant a shifts location without altering spread, whereas multiplying by b scales spread by $|b|$: $s_y = |b|\, s_x$ and $s_y^2 = b^2 s_x^2$.

Coefficient of Variation (CV), Skewness, and Kurtosis

Standard deviation cannot directly compare volatility across datasets measured in different units or vastly different scales (for example, comparing a A$45 share against a A$4.50 share, or comparing AUD revenue variability against USD unit cost variability). The **coefficient of variation (CV)** expresses sample standard deviation as a percentage of the sample mean (for ratio-scaled data with $\bar{x} > 0$):

$$\text{CV} = \left(\frac{s}{\bar{x}}\right) \times 100\%$$

Where: CV is the dimensionless coefficient of variation (in %), s is the sample standard deviation, and $\bar{x}$ is the sample mean.

DegreeLive

To quantify shape beyond location and spread, analysts compute higher-order standardised moments. **Sample skewness** measures asymmetry around the mean: positive values indicate a longer right tail, negative values indicate a longer left tail, and zero indicates symmetry. **Excess kurtosis** measures tail heaviness relative to a normal bell curve: positive excess kurtosis (leptokurtic distributions, characteristic of daily ASX equity returns) warns risk managers that extreme crashes and rallies occur far more frequently than a normal model predicts.

Worked Numerical Example 3: Dispersion, IQR, and Outliers in Melbourne Hospitality Wages

An industrial relations researcher samples $n = 10$ casual hourly wage rates (in A$) across independent Melbourne hospitality venues. The ordered observations are:

28.50, 29.00, 30.00, 31.00, 31.50, 32.50, 33.00, 34.50, 36.00, 44.00

Step 1: Compute the sample mean ($\bar{x}$).

$$\bar{x} = \frac{28.50 + 29.00 + 30.00 + 31.00 + 31.50 + 32.50 + 33.00 + 34.50 + 36.00 + 44.00}{10} = \frac{330.00}{10} = 33.00$$

Where: 330.00 is the sum of the ten hourly wages (A$) and $n = 10$.

Step 2: Compute squared deviations, sample variance (s^2), standard deviation (s), and CV.

Observation (x_i)	Deviation ($x_i - 33.00$)	Squared Deviation ($(x_i - 33.00)^2$)
28.50	-4.50	20.25
29.00	-4.00	16.00
30.00	-3.00	9.00
31.00	-2.00	4.00
31.50	-1.50	2.25
32.50	-0.50	0.25
33.00	0.00	0.00
34.50	+1.50	2.25
36.00	+3.00	9.00
44.00	+11.00	121.00
Sum = 330.00	**Sum = 0.00**	**SS = 184.00**

$$s^2 = \frac{184.00}{10 - 1} = \frac{184.00}{9} = 20.4444 \text{ AUD}^2 \quad \Longrightarrow \quad s = \sqrt{20.4444} = \text{AUD } 4.52$$

$$\text{CV} = \left(\frac{4.5216}{33.00}\right) \times 100\% = 13.70\%$$

Where: 184.00 is the sum of squared deviations, $n - 1 = 9$ is the degrees of freedom, s^2 is sample variance, s is sample standard deviation, and CV is the coefficient of variation.

DegreeLive

Step 3: Compute the five-number summary, IQR, and outlier fences. With $n = 10$ (even), the median Q_2 averages the 5th and 6th ordered values: $(31.50 + 32.50)/2 = \text{AUD } 32.00$. Using Tukey's halves: the lower half contains the first 5 values $\{28.50, 29.00, 30.00, 31.00, 31.50\}$, whose median is $Q_1 = \text{AUD } 30.00$; the upper half contains the last 5 values $\{32.50, 33.00, 34.50, 36.00, 44.00\}$, whose median is $Q_3 = \text{AUD } 34.50$. Therefore:

$$\text{IQR} = Q_3 - Q_1 = 34.50 - 30.00 = \text{AUD } 4.50$$

$$\text{Lower Fence} = 30.00 - 1.5(4.50) = \text{AUD } 23.25 \qquad \text{Upper Fence} = 34.50 + 1.5(4.50) = \text{AUD } 41.25$$

Where: $Q_1 = 30.00$ and $Q_3 = 34.50$ are the lower and upper quartiles. Because A$44.00 exceeds the upper fence of A$41.25 (perhaps reflecting a specialist supervisory or public-holiday penalty rate), A$44.00 is flagged as a statistical upper outlier. Note how removing or accounting for this single outlier explains why the mean (A$33.00) sits above the median (A$32.00) and why s (A$4.52) is inflated relative to the robust IQR (A$4.50).

1.7 Empirical Rule, Chebyshev's Theorem, and Standardisation (z-Scores)

Knowing a dataset's mean and standard deviation allows analysts to bound the proportion of observations lying within k standard deviations of the mean, and to convert raw figures into dimensionless standardised scores.

Chebyshev's Theorem versus the Empirical Rule

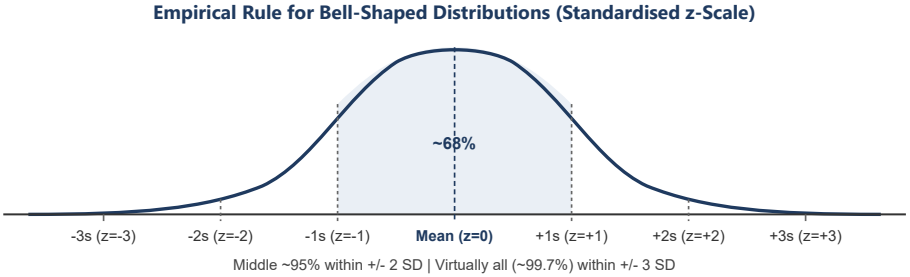
Chebyshev's Theorem applies universally to *any* dataset regardless of distributional shape (skewed, bimodal, or uniform). For any real constant $k > 1$, the proportion of observations lying within k standard deviations of the mean is at least:

$$P\left(|X - \mu| < k\sigma\right) \geq 1 - \frac{1}{k^2}$$

Where: $k > 1$ is the number of standard deviations from the mean, μ is the mean, and σ is the standard deviation. Setting $k = 2$ guarantees that at least $1 - 1/4 = 75\%$ of observations lie within $\mu \pm 2\sigma$; setting $k = 3$ guarantees at least $1 - 1/9 = 88.89\%$ lie within $\mu \pm 3\sigma$.

When a dataset is approximately symmetric and bell-shaped (normal), Chebyshev's conservative lower bounds sharpen dramatically into the **Empirical Rule (68–95–99.7 Rule)**:

- Approximately **68.27%** of observations lie within $\mu \pm 1\sigma$ (corresponding to $z \in [-1, +1]$).
- Approximately **95.45%** (commonly cited as 95%) of observations lie within $\mu \pm 2\sigma$ ($z \in [-2, +2]$).
- Approximately **99.73%** of observations lie within $\mu \pm 3\sigma$ ($z \in [-3, +3]$).



Standardisation (z-Scores)

To compare observations drawn from different units, scales, or benchmarks, analysts convert each raw value x_i into a dimensionless **standardised score (z-score)** representing the exact signed number of standard deviations by which x_i lies above or below the mean:

$$z_i = \frac{x_i - \bar{x}}{s} \quad \text{(sample)} \quad \text{and} \quad z_i = \frac{x_i - \mu}{\sigma} \quad \text{(population)}$$

Where: z_i is the standardised z-score for observation i, x_i is the raw observation, $\bar{x}$ (or μ) is the sample (or population) mean, and s (or σ) is the sample (or population) standard deviation.

Standardisation is an affine linear transformation ($a = -\bar{x}/s$, $b = 1/s$). Consequently, for any dataset with $s > 0$, the resulting z-scores always have a mean of exactly zero ($\bar{z} = 0$) and a standard deviation of exactly one ($s_z = 1$), while preserving the exact relative distances and skewness of the original distribution. In practice, observations satisfying $|z_i| > 3$ are conventionally flagged as extreme anomalies for data-quality auditing, and composite business indices (such as ESG scorecards or multi-criteria branch rankings) average z-scores across heterogeneous metrics so that no single large-scale metric dominates the index.

Worked Numerical Example 4: Cross-Market Performance Standardisation

Part A (Sales Manager Evaluation): A national Australian logistics firm evaluates two regional managers for an annual bonus. The Sydney manager generated A$685,000 in quarterly revenue in a New South Wales market averaging $\bar{x}_S = \text{AUD } 610,000$ with $s_S = \text{AUD } 50,000$. The Perth manager closed 4,820 freight contracts in a Western Australian market averaging $\bar{x}_P = 4,100$ contracts with $s_P = 600$ contracts. Which manager performed further above regional peers?

$$z_{\text{Sydney}} = \frac{685,000 - 610,000}{50,000} = \frac{75,000}{50,000} = +1.50$$

$$z_{\text{Perth}} = \frac{4,820 - 4,100}{600} = \frac{720}{600} = +1.20$$

Where: z_{Sydney} and z_{Perth} are the dimensionless standardised performance scores relative to each region's peer distribution. Although the raw metrics (A$ versus contract counts) are incomparable directly, standardisation shows that the Sydney manager sits 1.50 standard deviations above the NSW average compared with +1.20 standard deviations for the Perth manager, indicating stronger relative outperformance in Sydney.

Part B (Chebyshev versus Empirical Bounds): Customer call-handling times at an Australian superannuation call centre have mean $\mu = 240$ seconds and standard deviation $\sigma = 30$ seconds. Management investigates the interval $[180, 300]$ seconds, which corresponds to $k = (300 - 240)/30 = 2$ standard deviations around the mean. Without assuming normality, Chebyshev's theorem guarantees that at least $1 - 1/2^2 = 75\%$ of all calls finish between 180 and 300 seconds. If historical histograms confirm that call-handling times are approximately bell-shaped, the Empirical Rule refines that estimate to roughly 95% of all calls, with approximately 2.5% exceeding 300 seconds.

Degreelive