ACROSS MARKETS AND BORDERS

Enterprise, Trade and Leadership with Purpose and Integrity

PART I | FOUNDATIONS OF TRUTHFUL BUSINESS INQUIRY

CHAPTER 4

DATA ANALYTICS AND BUSINESS STATISTICS

From Data Quality to Statistical Evidence

Sixbert SANGWA

Department of International Business and Trade
African Leadership University, Kigali, Rwanda

Open Christian Press

2026

CHAPTER 4 | DATA ANALYTICS AND BUSINESS STATISTICS

From Data Quality to Statistical Evidence

First edition, 2026

Chapter DOI: https://doi.org/10.65655/ocp.m23.c63

Book DOI: https://doi.org/10.65655/ocp.m23

Author: Sixbert SANGWA

Affiliation: Department of International Business and Trade, African Leadership University, Kigali, Rwanda

Corresponding author: ssangwa@alueducation.com

Publisher: Open Christian Press, a Publishing Imprint of Open Christian University and Center for Faith and Work.

Publisher website: https://press.openchristian.education/

Publisher correspondence: editor@openchristian.education

Suggested citation. Sangwa, S. (2026). Data analytics and business statistics: From data quality to statistical evidence. In Across markets and borders: Enterprise, trade and leadership with purpose and integrity. Open Christian Press. https://doi.org/10.65655/ocp.m23.c63

Abstract

Data analytics turns observations into evidence, but statistical technique cannot rescue data whose meaning, quality or provenance is misunderstood. This chapter develops the quantitative foundation needed for responsible business and trade analysis. It follows the data lifecycle from business question, source and governance through cleaning, exploratory analysis and visualization, then builds probability, sampling distributions, confidence intervals, hypothesis testing, t-tests, chi-square tests, analysis of variance, correlation, simple and multiple regression, and introductory time-series forecasting. Worked examples connect formulas to managerial interpretation and repeatedly distinguish statistical significance from practical importance, association from causation, and model fit from truth. Rwanda's 2026 consumer-price data and World Bank Enterprise Survey metadata provide African applications, while large-scale online experimentation illustrates global statistical practice and the risks of multiple testing. Christian moral reasoning treats numerical analysis as accountable truth-telling: analysts must not manufacture certainty, hide inconvenient results, violate privacy or use technically valid summaries to mislead decision makers.

Keywords: business statistics; data analytics; data quality; descriptive statistics; probability; statistical inference; hypothesis testing; regression; time series; statistical ethics

Learning Outcomes

1	Define a business analytics question precisely and distinguish the decision question, statistical estimand, unit of analysis, variable types and population to which a quantitative claim refers.
2	Evaluate data sources and the data lifecycle using quality dimensions such as accuracy, completeness, consistency, timeliness, validity, uniqueness, relevance, provenance and fitness for purpose.
3	Create an auditable analytic dataset by applying disciplined rules for structure, data dictionaries, validation, missing data, outliers, transformation, version control and reproducible documentation.
4	Calculate and interpret frequencies, proportions, means, medians, quantiles, variance, standard deviation, coefficients of variation and other descriptive summaries, selecting measures suited to the distribution and decision.
5	Design and critique statistical visualizations and basic dashboards, recognizing misleading scales, denominator choices, aggregation effects, uncertainty omissions and accessibility failures.
6	Apply probability rules, conditional probability, Bayes reasoning, common probability distributions, expected value, variance, sampling distributions, standard errors and the central limit theorem to business uncertainty.
7	Construct and correctly interpret confidence intervals and hypothesis tests, including p-values, significance levels, Type I and Type II errors, statistical power and the difference between statistical and practical significance.
8	Select, compute and interpret one-sample, paired and independent-samples t-tests, chi-square tests and one-way ANOVA while checking the assumptions and limitations that govern each method.
9	Analyze association using covariance, Pearson and Spearman correlation without converting association into an unsupported causal claim.
10	Estimate and interpret simple and multiple linear regression models, diagnose major assumption problems, distinguish conditional association from causation, and communicate effect sizes and uncertainty.
11	Decompose introductory business time series into level, trend, seasonality and irregular variation, construct simple benchmark forecasts and evaluate forecast error without pre-empting advanced predictive analytics in Chapter 5.
12	Apply Christian moral reasoning and professional statistical ethics to privacy, bias, selective reporting, reproducibility, data visualization, uncertainty and communication to non-technical decision makers.

Concept Map and Chapter Roadmap

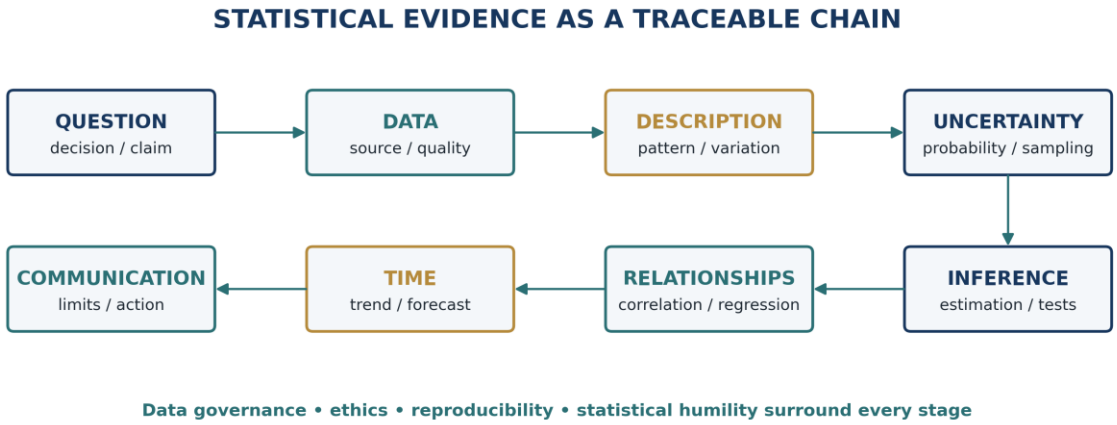

Figure 4.1. Statistical evidence as a traceable chain from question to communication.
Source: Author's synthesis; chapter architecture follows the approved Book Blueprint.

The chapter moves from provenance to probability. Sections 4.1-4.6 establish what a quantitative dataset is, how it is governed, and why cleaning is an inferential rather than merely clerical task. Sections 4.7-4.9 develop descriptive statistics, exploratory analysis and visualization. Sections 4.10-4.14 build probability, distributions, sampling uncertainty, confidence intervals and hypothesis testing. Sections 4.15-4.18 apply those foundations to t-tests, chi-square tests, ANOVA, effect sizes, multiple testing and power. Sections 4.19-4.21 develop correlation, linear regression and introductory time-series forecasting. The final sections integrate African and global cases, moral reasoning, a professional evidence protocol, common errors, applied exercises and the hand-off to Chapter 5. The chapter therefore owns statistical evidence, not advanced predictive machine learning, optimization, simulation, scenario design or managerial decision architecture.

Opening Case: Rwanda's 2026 Inflation Data and the Difference Between a Number and Evidence

In August 2026, the National Institute of Statistics of Rwanda reported that Rwanda's Consumer Price Index had increased by 14.5% in July 2026 compared with July 2025, up from 13.6% year on year in June. Within the July release, transport prices were reported 24.2% higher than a year earlier, energy 44.5% higher, food and non-alcoholic beverages 13.1% higher, imported products 10.7% higher, and local products 15.8% higher (National Institute of Statistics of Rwanda [NISR], 2026b). For a Kigali retailer, manufacturer, investor or employer, those numbers are commercially consequential. They can influence pricing, wage discussions, inventory policy, working-capital needs, procurement contracts and forecasts of customer purchasing power.

Yet a competent analyst cannot move directly from the headline '14.5%' to a managerial conclusion. The CPI is an index constructed from a defined basket and weighting system; it is not the percentage by which every household's or firm's costs increased. The annual rate compares two index levels twelve months apart; it is not the same as the July month-on-month change. The energy component can rise much faster than the all-items index without implying that energy accounts for 44.5% of the average basket. A business whose cost structure is unusually transport-intensive can face effective inflation well above the headline rate, while another firm with long-term contracts or different inputs can experience much less. Statistical evidence begins by asking what the number actually measures.

The time path also matters. NISR's monthly releases reported year-on-year headline inflation of 8.9% in January, 9.2% in February, 9.2% in March, 13.0% in April, 12.9% in May, 13.6% in June and 14.5% in July 2026 (NISR, 2026a). A graph makes the acceleration visible, but the graph does not explain its causes. A regression can summarize association between inflation and another variable, but the coefficient will not become causal simply because software prints a p-value. A forecast can extend the series, but a short sequence exposed to policy, exchange-rate, harvest, energy and external shocks carries substantial uncertainty. The correct professional question is therefore not 'What does the software say?' but 'What claim is justified by these data, under these assumptions, and with what uncertainty?'

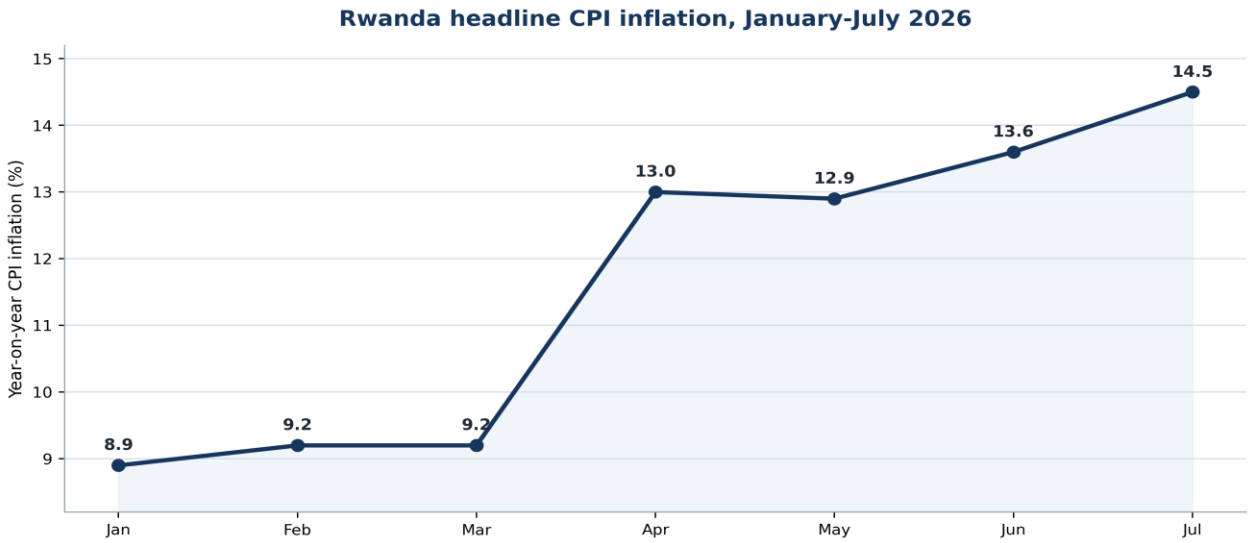

Figure 4.2. Rwanda headline CPI inflation, January-July 2026.

Source: National Institute of Statistics of Rwanda (2026a). Values are year-on-year headline CPI inflation rates, not month-on-month changes.

MANAGERIAL INSIGHT | A STATISTIC IS NOT SELF-INTERPRETING

A number becomes evidence only after its definition, denominator, time basis, data-generating process, uncertainty and relevance to the decision are understood. The discipline applies equally to inflation, sales conversion, employee turnover, defect rates, market shares and investment returns.

The opening case gives the chapter its central problem: How can business professionals transform quantitative observations into statistical evidence without confusing calculation with knowledge, precision with certainty, or association with causation?

4.1 Why Data Analytics and Business Statistics Matter

Chapter 3 established how research questions, populations, samples, measures and designs create the conditions under which evidence can be generated. Chapter 4 begins after those design choices have produced quantitative observations. It asks how those observations should be structured, summarized, visualized, compared and modeled so that the resulting claims are defensible. The hand-off matters because statistics cannot repair a sample that never represented the intended population, a variable that never measured the construct, or a design that never created a credible comparison. Statistical analysis inherits the strengths and weaknesses of the data-generating process described in Chapter 3.

Business statistics is the disciplined use of quantitative methods to describe variation, quantify uncertainty, estimate population features and assess relationships relevant to business decisions. Data analytics is broader in everyday professional usage, often including data acquisition, cleaning, exploration, visualization and the transformation of raw records into usable evidence. In this chapter the two are integrated at the foundational level. The aim is not to create a catalogue of software commands but to develop statistical judgment: knowing what a method assumes, what question it answers, how uncertainty should be represented, and when a technically available method should not be used.

The chapter's approved boundary is important. Descriptive and inferential statistics, classical regression and introductory forecasting belong here. Chapter 5 moves from evidence to structured managerial action and owns predictive machine learning, classification, clustering, optimization, simulation, scenario models, advanced forecasting for planning, decision trees, multi-criteria decision analysis and the governance of AI-supported decisions. Chapter 4 may show why a regression coefficient is uncertain; Chapter 5 asks how a decision system should use predictions and trade-offs. This division prevents the common mistake of treating 'analytics' as one undifferentiated toolbox.

4.1.1 Statistical reasoning: from description to inference

Modern business statistics developed through complementary traditions rather than a single method. Student (1908) showed how small-sample uncertainty about a mean could be handled when population variance was unknown; Fisher (1925) systematized estimation, experimental comparison and analysis of variance; and Tukey (1977) later insisted that

exploratory examination of data should precede rigid modeling. Contemporary analytics adds data engineering, high-volume computation and reproducible workflows, but the underlying epistemic problem is unchanged: separate systematic pattern from variation while making the assumptions behind that separation visible.

Three layers of reasoning should therefore remain distinct. Description summarizes what was observed. Statistical inference uses a sampling or probability model to estimate beyond the observed sample and quantify uncertainty. Decision analysis asks what action should follow once objectives, costs, risks and moral constraints are considered. Chapter 4 owns the first two layers; Chapter 5 develops the third. Confusing them is a recurring professional error because an accurate description need not generalize, and a statistically defensible estimate does not by itself determine what an organization ought to do.

EVIDENCE CHECK | STATISTICAL SOPHISTICATION IS NOT SOFTWARE SOPHISTICATION

A model with many variables, an interactive dashboard or a generative-AI explanation can look advanced while resting on weak definitions, duplicated records or biased selection. The hierarchy should run in the opposite direction: question first, data provenance second, quality third, statistical method fourth, communication fifth.

4.2 From a Business Question to an Analytic Dataset

4.2.1 Decision question, statistical question and estimand

A decision question asks what an organization should do. A statistical question asks what quantitative feature must be learned before that decision can be informed. The estimand is the precise quantity the analysis seeks to estimate. Suppose a regional distributor asks whether delivery performance has deteriorated. ‘Should we add another warehouse?’ is a decision question. ‘What was the mean on-time-delivery rate during the last quarter, and how did it differ from the previous quarter for comparable routes?’ is a statistical question. The corresponding estimand might be a difference in population proportions under a specified definition of ‘on time.’

This distinction disciplines analysis. Without an estimand, analysts can search a dataset until something interesting appears. With an estimand, variable definitions, denominators, time windows and comparison groups become explicit before results are known. This is especially important in high-dimensional business data because the number of possible summaries and subgroup comparisons can be enormous.

4.2.2 Unit of analysis, observation and aggregation

The unit of analysis is the entity about which the claim is made: customer, order, shipment, employee, branch, firm, country or month. The unit of observation is the level at which records are stored. A firm-level claim based on transaction-level records requires appropriate aggregation; an employee-level regression with repeated monthly observations requires recognition that records from the same employee are not independent. Chapter 3 introduced these distinctions in research design. Chapter 4 adds the statistical consequence: standard errors, averages and tests can be wrong when clustered or repeated observations are treated as independent copies.

Aggregation can also hide structure. A national average may conceal regional variation; a monthly average can conceal day-of-week peaks; a firm-wide conversion rate can reverse within customer segments. Simpson’s paradox describes situations in which an aggregate association differs from, or reverses, associations within groups because group composition changes. The practical lesson is not ‘never aggregate’ but ‘aggregate only at the level that answers the question, and inspect important stratification before interpreting the average.’

4.2.3 Variables, scales and analytical consequences

Chapter 3 addressed operationalization and measurement validity. Here the minimal reminder is that a variable’s statistical type constrains sensible analysis. Nominal categories have labels without an intrinsic ordering; ordinal categories have an ordering without assured equal spacing; interval measures have meaningful differences but no true zero; ratio measures support meaningful ratios. In practical analytics it is often more useful to distinguish categorical variables from quantitative variables and then record whether quantitative observations are discrete or continuous, while preserving the original measurement logic.

Table 4.1. Variable types and common statistical treatments

Variable type	Business example	Meaningful summaries	Typical cautions

Variable type	Business example	Meaningful summaries	Typical cautions
Nominal categorical	Payment channel: cash, card, mobile money	Counts, proportions, mode, cross-tabulation	Do not average numeric category codes.
Ordinal categorical	Service rating: poor to excellent	Counts, median/rank summaries, ordered plots	Distances between adjacent categories may not be equal.
Discrete quantitative	Number of defects or orders	Mean, variance, counts; count distributions	Strong skew and many zeros are common.
Continuous quantitative	Revenue, time, weight, price	Mean/median, SD/IQR, regression, intervals	Scale, skew, censoring and currency units matter.
Binary indicator	Default: yes/no	Proportion, risk difference, odds/risk models	Coding 0/1 does not make the variable continuous.
Time index	Daily sales, monthly inflation	Trend, seasonality, lags, forecast error	Ordering and autocorrelation violate ordinary independence assumptions.

Source/Note: Author's synthesis. Measurement validity and operationalization are developed in Chapter 3.

4.3 Data Sources, Structures and the Data Lifecycle

4.3.1 Transactional, administrative, survey, experimental and external data

Business datasets enter analysis through different institutional processes. Transactional systems record sales, payments, clicks, inventory movements or service events. Administrative systems record payroll, licensing, customs, tax, attendance or compliance events because an organization must operate, not primarily because a researcher designed a study. Surveys generate structured responses from sampled units. Experiments record outcomes under assigned conditions. External data may come from national statistical offices, central banks, market-data providers, geospatial sources, platform APIs or international institutions. Each source carries different strengths and blind spots.

Large volume does not remove selection. A complete database of customers is still a selected population if the question concerns all potential customers. Platform logs can record every click while missing people who never gained access to the platform. Administrative records can be precise about registered firms and silent about informal activity. National indicators can be authoritative for the population they were designed to represent but inappropriate for a firm-specific operational question. Provenance therefore precedes computation.

4.3.2 Structured, semi-structured and unstructured data

Structured data have a stable schema of rows, columns and field types. Semi-structured data such as JSON, XML or event logs contain labels and hierarchy without a fixed rectangular table. Unstructured data include documents, images, audio and free text. Modern firms often transform semi-structured and unstructured material into structured features before statistical analysis. That transformation itself creates measurement choices. A sentiment score extracted from customer text, for example, is not the text; it is a model-generated representation whose validity and error must be evaluated.

4.3.3 The lifecycle and data lineage

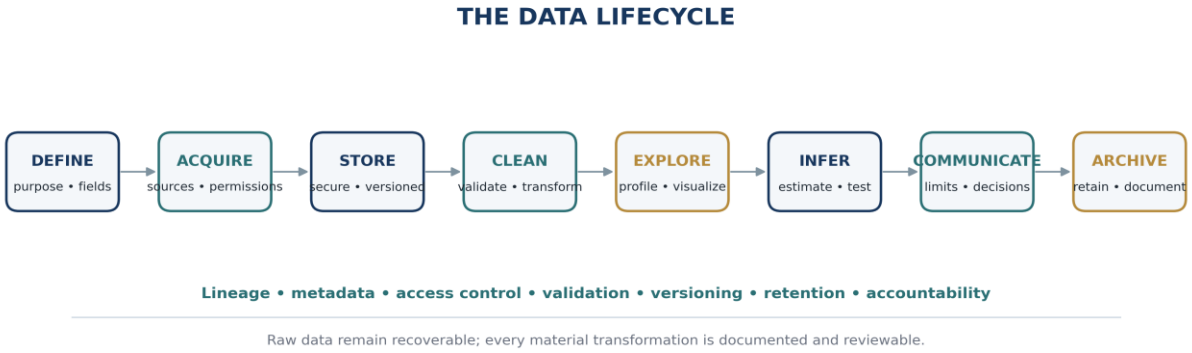

Figure 4.3. The data lifecycle from definition to accountable reuse.
Source: Author's synthesis, informed by data-management and reproducibility literature.

A defensible lifecycle begins with definition, not collection. Variables and identifiers are specified; sources and permissions are established; raw data are acquired without destructive modification; storage protects confidentiality and integrity; cleaning produces an analytically usable version; exploratory analysis checks structure and anomalies; inference answers pre-specified or clearly labeled exploratory questions; communication represents uncertainty; and archiving preserves the data dictionary, code, decisions and version history needed for later review. Lineage is the traceable record of where each analytical field came from and how it was transformed.

Wickham's (2014) tidy-data principle offers a useful structural discipline: each variable forms a column, each observation a row and each type of observational unit a table. Broman and Woo (2018) add spreadsheet practices that reduce error, such as putting one thing in a cell, using consistent date formats, avoiding information encoded only through color, maintaining a data dictionary and separating raw data from calculations. These are not aesthetic preferences. They make transformations inspectable and reduce silent ambiguity.

PRACTICE TOOL | THE RAW-PROCESSED-OUTPUT RULE

Maintain at least three logical layers. Raw data are immutable copies of received records. Processed data contain documented cleaning and derived variables. Outputs contain tables, figures and models generated from the processed layer. If analysts manually overwrite raw cells, the audit trail collapses and later results cannot be reproduced reliably.

4.4 Data Governance, Privacy, Security and Sovereignty

Data governance assigns authority and responsibility for data definitions, quality, access, retention, security, sharing and deletion. Statistics is therefore embedded in governance. An analyst cannot responsibly maximize analytical convenience by copying identifiable customer data into personal spreadsheets, uploading confidential records to an unapproved AI service, or retaining sensitive fields indefinitely 'in case they become useful.' Data minimization, role-based access, secure storage, documented purpose and retention rules reduce both legal exposure and moral risk.

Rwanda's Law No. 058/2021 relating to the protection of personal data and privacy applies to electronic and non-electronic processing and extends to certain controllers or processors outside Rwanda when they process personal data of data subjects located in Rwanda (Republic of Rwanda, 2021). The African Union Data Policy Framework seeks stronger and more harmonized governance across the continent while linking data use to rights, development and intra-African digital trade (African Union, 2022). Such frameworks can support interoperability and trust, but harmonization should not be treated as an automatically superior outcome. Analysts and policymakers must still examine proportionality, sovereignty, local institutional capacity, cross-border transfer risks, accountability and whether data concentration expands power without adequate redress.

The FAIR principles - findable, accessible, interoperable and reusable - were developed for scientific data stewardship and are often valuable for analytical workflows (Wilkinson et al., 2016). Accessibility in FAIR does not mean unrestricted public access. Sensitive data can be discoverable through metadata while remaining subject to lawful access controls. In business, the same distinction prevents a false choice between reproducibility and privacy:

reproducible work can preserve code, schemas, synthetic examples and controlled access without exposing personal records.

POLICY LENS | DATA VALUE DOES NOT ERASE DATA RIGHTS

Data can create commercial and public value while also increasing surveillance, discrimination or concentration of power. The professional standard is not to maximize data collection. It is to collect and retain what is justified by purpose, law, risk and legitimate human benefit, while preserving the ability to audit analytical claims.

4.5 Data Quality: Fitness for Statistical Purpose

4.5.1 Accuracy is necessary but not sufficient

Data quality is often reduced to accuracy, but an accurate value can still be unusable if it is incomplete, stale, inconsistently defined, duplicated, inaccessible or irrelevant to the question. Wang and Strong's (1996) influential consumer-oriented framework distinguished intrinsic, contextual, representational and accessibility dimensions, emphasizing that quality is partly task-dependent. A perfectly recorded 2022 exchange rate is not high-quality evidence for a transaction priced in August 2026. A complete customer dataset can be low quality if the customer identifier changes across systems and duplicates cannot be resolved.

Table 4.2. Core data-quality dimensions for business statistics

Dimension	Diagnostic question	Example failure	Statistical consequence
Accuracy	Does the value correspond to reality within acceptable error?	Invoice amount entered as 1,500 instead of 15,000	Biased totals and model estimates.
Completeness	Are required records and fields present?	Default outcomes missing for closed accounts	Biased estimates if missingness is systematic.
Consistency	Do definitions and codes agree across sources/time?	"SME" changes threshold across years	Invalid comparisons.
Timeliness	Is the data current enough for the decision?	Inventory file updated weekly in a volatile setting	Stale estimates and poor forecasts.
Validity	Does the value satisfy domain rules?	Negative delivery time; invalid country code	Signals entry, parsing or schema error.
Uniqueness	Is each real entity represented as intended?	Same customer stored under two IDs	Double counting.
Relevance	Does the field answer the stated question?	Likes used as a substitute for sales	Metric-substitution error.
Provenance/lineage	Can origin and transformation be traced?	Copied column with unknown formula	Results cannot be audited or reproduced.

Source/Note: Author's synthesis, drawing on Wang and Strong (1996), Wickham (2014), Broman and Woo (2018), and reproducibility principles.

4.5.2 Data profiling

Data profiling is the systematic inspection of a dataset before substantive modeling. At minimum, the analyst should inspect row and column counts, field types, valid ranges, unique values, duplicates, missingness rates, date coverage, distribution summaries, impossible combinations and identifier integrity. Profiling should be automated where possible so that later data refreshes are tested against the same rules. A one-time visual scan is inadequate for recurring operational analytics.

Quality thresholds should be linked to use. A small number of missing optional marketing fields may be tolerable for a revenue summary but unacceptable for a fairness audit if the missing field is required to examine differential treatment. A 1% duplicate rate might be immaterial for one estimate and catastrophic for fraud detection. 'Clean' is therefore not a universal state; it means documented fitness for a specified analytical purpose.

4.6 Cleaning and Transformation: Missing Values, Outliers and Documentation

4.6.1 Cleaning should preserve evidence, not beautify it

Cleaning is the controlled correction or representation of known data problems. It should not become a process of forcing observations to look normal, removing inconvenient cases or smoothing away genuine variation. Every material rule should be explainable: which records were changed, why, using what source of truth, and how the original can be recovered. The analyst should prefer reproducible transformations in code or documented queries to unrecorded manual edits.

4.6.2 Missing data are a mechanism problem

A missing value is not a zero and not automatically random. Rubin (1976) formalized distinctions that remain useful. Data are missing completely at random when missingness is unrelated to observed or unobserved values relevant to the analysis; missing at random when missingness can be explained by observed information under the model; and missing not at random when missingness depends on unobserved values even after conditioning on observed data. These labels are assumptions about the missingness process, not patterns that software can prove from the incomplete dataset alone.

Complete-case analysis is simple but can waste information and produce bias when excluded cases differ systematically. Single imputation with the overall mean preserves sample size but artificially reduces variance and weakens relationships. Multiple imputation, likelihood-based methods, weighting and sensitivity analysis are more principled under appropriate assumptions (Little & Rubin, 2019). At undergraduate level, the essential discipline is to report the amount and pattern of missingness, justify the handling method, and examine whether conclusions depend materially on plausible alternatives.

4.6.3 Outliers: error, rare event or important signal?

An outlier is an observation unusual relative to the rest of the data under some criterion. It can be a data-entry error, a unit mismatch, a legitimate extreme value, a structural break, fraud, a premium customer, a crisis event or evidence that a model is misspecified. Automatic deletion is therefore poor practice. Analysts should first verify provenance, units and context. If the value is genuine, robust summaries such as the median and interquartile range may describe the typical case better than the mean and standard deviation, while regression diagnostics can assess whether the observation is influential.

The familiar 1.5 × IQR boxplot rule is a flagging convention, not a universal law. In heavy-tailed revenue data, many legitimate observations can fall beyond the whiskers. Z-score rules similarly depend on distributional assumptions. Winsorizing or trimming can be justified for specific estimands, but the threshold must be pre-specified or transparently motivated and sensitivity results should be reported when conclusions depend on the treatment of extremes.

4.6.4 Transformation and standardization

Transformations can improve interpretability or model fit. Log transformation can reduce right skew and convert multiplicative relationships into additive ones, although zero and negative values require special treatment. Standardization expresses values in standard-deviation units, z = (x - mean)/SD, which helps compare coefficients or variables measured on different scales. Min-max normalization rescales values to a fixed interval but is sensitive to extremes. Transformations alter the scale on which effects are interpreted; they should therefore be chosen for a reason rather than applied mechanically.

Table 4.3. Data-problem decision guide

Problem	First diagnostic	Potential response	What not to do
Missing values	Pattern by variable, group and outcome	Document; model/weight/impute where justified; sensitivity analysis	Replace all missing values with zero or the mean.
Duplicate records	Identifier and event logic	Resolve exact/near duplicates using source-of-truth rules	Drop duplicates solely because rows look similar.
Outliers	Verify units, source and context	Correct errors; retain genuine extremes; use robust analysis/sensitivity	Delete values simply because they change the result.

Problem	First diagnostic	Potential response	What not to do
Inconsistent categories	Compare codebooks and historical definitions	Map to controlled vocabulary with documented crosswalk	Silently merge categories with different meanings.
Impossible values	Domain/range validation	Correct from source or mark missing with reason code	Force values into range without provenance.
Skewed scale	Histogram/quantiles; substantive scale meaning	Use median/IQR or justified transformation	Assume normality because sample is large.

Source/Note: Author's synthesis. Missing-data methods follow Rubin (1976) and Little and Rubin (2019).

ETHICAL DILEMMA | IS DELETING "MESSY" DATA EVER NEUTRAL?

Suppose a lender's rejected applications contain more missing income fields than approved applications. Dropping incomplete rows may yield a clean regression while systematically removing applicants who already faced exclusion. Cleaning decisions can redistribute visibility. The analyst must ask whose records become statistically absent and whether that absence changes the claim.

4.7 Descriptive Statistics: Center, Spread, Position and Shape

4.7.1 Frequencies, proportions and rates

Categorical data are commonly summarized with counts and proportions. A proportion is a count divided by a clearly defined denominator; a rate usually incorporates exposure to time or opportunity. 'Ten defects' is uninterpretable without production volume. Ten defects in 100 units is different from ten in 100,000. Business dashboards frequently fail not because the numerator is wrong but because the denominator changes silently across periods or groups.

$$\hat{p} = x / n$$

Define the numerator x, the relevant denominator n and all exclusions. A rate may additionally divide by time, transactions or another exposure measure.

4.7.2 Mean, median and mode

The arithmetic mean uses every value and is the natural center for many additive quantities, but it can be pulled strongly by extreme observations. The median is the 50th percentile and is robust to extremes, making it useful for skewed income, transaction-value or delivery-time distributions. The mode is the most frequent value or category and is most useful for discrete or categorical variables. No measure of center is universally best; the choice should match the distribution and the business question.

$$\bar{x} = (1/n) \sum x_i \quad | \quad \bar{x}w = \sum(w_i x_i) / \sum w_i$$

Weights must have a documented interpretation, such as survey selection weights, exposure weights or portfolio weights.

WORKED EXAMPLE | WHY WEIGHTING CAN CHANGE A BUSINESS CONCLUSION

Suppose three market segments have average order values of US$20, US$45 and US$100 and represent 70%, 25% and 5% of customers. The unweighted mean of the three segment means is US$55. The customer-weighted mean is 0.70(20) + 0.25(45) + 0.05(100) = US$30.25. Averaging averages without appropriate weights can badly misrepresent the overall customer base.

4.7.3 Variance, standard deviation and interquartile range

Measures of spread describe how observations vary around the center. The sample variance averages squared deviations using n - 1 in the denominator to estimate population variance under standard random-sampling assumptions. The standard deviation is the square root of variance and therefore returns to the original unit. The interquartile range, Q3 - Q1, measures the middle 50% and is resistant to extremes. The range is easy to compute but unstable because it depends only on the minimum and maximum.

$$s^2 = \sum(x_i - \bar{x})^2 / (n - 1) \quad | \quad s = \sqrt{s^2}$$

For positive ratio-scale variables with a meaningful zero, CV = s/x̄ can compare relative dispersion.

4.7.4 Percentiles, standardized scores and shape

Percentiles locate an observation relative to a distribution. A 90th-percentile delivery time means 90% of observed deliveries were no slower than that value under the chosen sample and period; it does not mean the delivery was '90% good.' Standardized z-scores measure distance from the mean in standard-deviation units. Skewness describes asymmetry, while kurtosis-related measures describe tail behavior. These shape descriptors can be useful, but graphs and quantiles often communicate distributional structure more directly than a single shape coefficient.

$$z = (x - \bar{x}) / s$$

A z-score is a relative position on the observed scale; it is not automatically a probability unless a distributional model is specified.

Table 4.4. Choosing descriptive summaries by distribution and purpose

Situation	Center	Spread	Helpful visual	Interpretive caution
Approximately symmetric continuous data	Mean	SD	Histogram / density / boxplot	Check for multimodality and outliers.
Strongly right-skewed monetary data	Median	IQR / percentiles	Histogram on sensible scale / boxplot	Mean may still matter for totals and budgeting.
Categorical data	Mode / proportions	Not SD	Bar chart	Codes are labels, not quantities.
Service-level tail performance	Median plus high percentile	IQR plus P90/P95	ECDF / boxplot	Average can conceal severe tail delays.
Comparing relative variability	Mean plus CV when appropriate	CV	Grouped dot/interval plot	CV is inappropriate near zero or for interval scales.

Source/Note: Author's synthesis.

4.8 Exploratory Data Analysis: Seeing Structure Before Testing Claims

Exploratory data analysis (EDA) uses numerical and graphical summaries to discover structure, anomalies and plausible relationships before or alongside formal modeling. Tukey (1977) emphasized the value of allowing data to reveal features that rigid confirmatory procedures can miss. EDA is not permission to present discovered patterns as if they had been pre-specified hypotheses. Its strength is diagnostic and generative: it helps analysts understand distributions, identify errors, formulate better models and ask where a result may differ across groups.

Useful EDA begins univariately and then adds relationships. For each important quantitative variable, inspect missingness, range, center, spread, quantiles and shape. For categorical variables, inspect levels and proportions. Then examine grouped summaries, cross-tabulations, scatterplots and time plots. Plot before modeling whenever feasible. Anscombe's (1973) quartet famously shows different datasets with nearly identical means, variances, correlations and fitted lines but radically different graphical structure. The lesson remains fundamental: summary statistics can conceal nonlinearity, outliers and clustering.

4.8.1 Cross-tabulation and conditional summaries

A cross-tabulation shows joint frequencies for two categorical variables and can be converted to row or column percentages depending on the question. The denominator should follow the direction of inference. If the question is 'Among exporters, what share use documentary credit?' use exporter counts as the denominator. If the question is 'Among documentary-credit users, what share are exporters?' the denominator changes. Tables that mix denominators can create persuasive-looking but logically incoherent comparisons.

For quantitative outcomes, grouped summaries should include sample size and uncertainty where possible. Reporting only a group mean can hide a tiny subgroup, high variability or an influential observation. Where data are

clustered by branch, country or repeated customer, analysts should preserve that structure for inferential methods rather than flattening it during EDA.

EVIDENCE CHECK | EXPLORE FREELY, LABEL HONESTLY

It is legitimate to discover an unexpected pattern and then investigate it. It is not legitimate to conduct dozens of exploratory comparisons and report the most favorable one as though it had been the sole planned test. Distinguish exploratory from confirmatory analysis and account for multiplicity when inference follows broad searching.

4.9 Data Visualization, Dashboards and Responsible Statistical Storytelling

4.9.1 Match the visual encoding to the comparison

Visualization converts statistical structure into perceptual structure. Cleveland and McGill's (1984) experiments showed that people decode some graphical elements more accurately than others, with position along a common scale generally supporting more precise comparison than area or volume. Practical chart choice should therefore begin with the comparison. Bars or dot plots support category comparisons; lines support ordered time; scatterplots support relationships; histograms and boxplots support distributions; interval plots support estimates with uncertainty.

Table 4.5. Visualization selection guide

Analytical purpose	Usually effective	Often weak or risky	Key integrity question
Compare categories	Sorted bar or dot plot	3-D bars, decorative pictograms	Does the axis and ordering permit fair comparison?
Show change over time	Line chart	Unordered bars for long series	Are intervals equally spaced and breaks disclosed?
Show a distribution	Histogram, boxplot, ECDF	Pie chart of continuous values	Are bin widths or truncation hiding structure?
Show relationship	Scatterplot, fitted line with uncertainty	Dual-axis line charts implying association	Do scales create an artificial pattern?
Show composition	100% bars, small multiples	Many-slice pies	Do percentages sum to a documented whole?
Show estimates	Point/interval plot	Stars for significance only	Are effect size and uncertainty visible?

Source/Note: Author's synthesis, informed by Cleveland and McGill (1984) and statistical-graphics principles.

4.9.2 Common visual distortions

A chart can be numerically correct and still misleading. Truncating the vertical axis can exaggerate small differences in bar charts; forcing a zero baseline on a line chart can sometimes hide meaningful variation. The correct choice depends on the task and should be disclosed. Dual axes can manufacture visual correlation by arbitrary scaling. Unequal bin widths can distort histograms if area is not handled correctly. Cumulative totals can look impressive even when current performance is deteriorating. Percent changes can explode when the baseline is near zero. A responsible visual makes the denominator, time basis and units visible and uses annotations to clarify genuine events rather than decorate the page.

4.9.3 Dashboards: monitoring, not automatic decision-making

A dashboard brings selected indicators into a recurring monitoring view. Chapter 5 develops business intelligence systems, KPIs and decision architecture; Chapter 4's concern is statistical quality. Every dashboard metric should have a definition, owner, refresh frequency, source, denominator, quality check and escalation rule. A red-green status display without uncertainty can create false certainty when metrics are noisy. Control limits, confidence bands or sample sizes can help users distinguish routine variation from changes worth investigation.

Responsible data storytelling connects evidence to a question without turning a narrative into causal proof. Chapter 2 developed communication principles. Here the statistical extension is to show magnitude, comparison, denominator and uncertainty before emphasizing a conclusion. If a graph is filtered, rebased, normalized or seasonally adjusted, those transformations should be disclosed. A non-technical audience deserves less jargon, not less truth.

MANAGERIAL INSIGHT | THE DASHBOARD SHOULD MAKE BAD NEWS VISIBLE

A dashboard designed mainly to reassure senior leaders is a reporting risk. Statistical governance should protect inconvenient variance, failed targets and unexpected subgroup results from being silently suppressed by formatting choices or metric selection.

4.10 Probability as a Language of Uncertainty

Probability provides a formal language for uncertainty. It does not eliminate uncertainty; it organizes it. In business, uncertain events include whether a customer defaults, a shipment is delayed, a campaign converts, a machine fails, a competitor enters, a currency moves beyond a threshold or a project exceeds budget. Statistical inference depends on probability because samples vary even when drawn from the same population, and forecasts vary because future outcomes are not known.

4.10.1 Events, complements and addition rules

An event is a set of outcomes of interest. If A denotes 'a shipment is delayed,' its complement A^c denotes 'the shipment is not delayed.' Probabilities lie between 0 and 1 and the probabilities of mutually exclusive, collectively exhaustive outcomes sum to 1. The addition rule distinguishes events that can occur together from those that cannot.

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

The intersection term prevents double counting when A and B can occur together.

4.10.2 Conditional probability and independence

Conditional probability asks how the probability of one event changes when another event is known. The probability that a shipment is delayed may be 8% overall but 22% given a severe-weather alert. Conditional probability is written $P(A|B)$. Events A and B are independent when knowing B does not change the probability of A, so $P(A|B) = P(A)$. Independence should never be assumed simply because two variables have different names; operational processes often create dependence.

$$P(A \mid B) = P(A \cap B) / P(B), \text{ for } P(B) > 0$$

Equivalently, $P(A \cap B) = P(A \mid B)P(B)$.

4.10.3 Bayes' theorem and base rates

Bayes' theorem reverses a conditional probability by combining new evidence with a prior or base rate. This is especially important in fraud, quality control and screening, where rare events can produce many false alarms even when a detection system is highly sensitive.

$$P(A \mid B) = P(B \mid A)P(A) / P(B)$$

For binary A: $P(B) = P(B \mid A)P(A) + P(B \mid A^c)P(A^c)$.

WORKED EXAMPLE | WHY A 90% ACCURATE ALERT CAN STILL BE MOSTLY WRONG

Suppose 1% of transactions are fraudulent. A detection rule flags 90% of frauds and falsely flags 5% of legitimate transactions. Among 10,000 transactions, about 100 are fraudulent and 90 are correctly flagged. Of the 9,900 legitimate transactions, about 495 are falsely flagged. A flagged transaction therefore has a fraud probability of about 90/(90+495) = 15.4%, not 90%. The base rate matters.

This example also illustrates why 'accuracy' can be a poor metric in imbalanced problems. A system that labels every transaction legitimate would be 99% accurate in this scenario while detecting no fraud. Chapter 5 develops classification metrics and predictive decision rules; Chapter 4's role is to establish the probability logic behind them.

4.11 Random Variables and Common Probability Distributions

4.11.1 Discrete and continuous random variables

A random variable assigns a numerical value to uncertain outcomes. Discrete random variables take countable values, such as number of defaults in a portfolio. Continuous random variables can take values across an interval, such as delivery time or transaction value. A probability distribution specifies how probability is allocated across possible values. Its expected value is the probability-weighted long-run average under the model; its variance quantifies dispersion around that expected value.

$$E(X) = \sum xP(X = x)$$

For a continuous variable, expectation is defined by integration; Var(X) = E[(X - E(X))²].

4.11.2 Bernoulli and binomial distributions

A Bernoulli variable has two outcomes commonly coded 1 and 0, with success probability p. If n independent Bernoulli trials share the same p, the number of successes follows a binomial distribution. Binomial models can approximate situations such as conversions among independent visitors or defect counts among independently produced units, but the assumptions should be checked. Customer outcomes can be correlated through campaigns, geography or shared shocks, and success probabilities can differ across individuals.

$$\text{If } X \sim \text{Binomial}(n, p): \quad E(X) = np \quad | \quad Var(X) = np(1 - p)$$

The basic binomial model assumes fixed n, two outcomes per trial, common p and independence across trials.

4.11.3 Poisson distribution

The Poisson distribution models counts of events occurring over a specified exposure when events are approximately independent and occur at a stable average rate. Examples include arrivals, defects or claims per interval. Its mean and variance are both λ under the basic model. Real business counts often show overdispersion, clustering or excess zeros, in which case the simple Poisson model understates variability.

4.11.4 Normal, t, chi-square and F distributions

The normal distribution is continuous, symmetric and defined by mean μ and variance σ². Its importance comes partly from empirical usefulness for many aggregated measurements and partly from the central limit theorem. The t distribution resembles the normal distribution but has heavier tails, reflecting uncertainty when a population variance is estimated from the sample. Chi-square distributions arise in variance and categorical-frequency inference. F distributions arise from ratios of variance estimates and underpin ANOVA. These distributions are not arbitrary software options; they encode sampling assumptions used to calculate uncertainty.

Table 4.6. Common probability distributions in foundational business statistics

Distribution	Typical variable	Parameters	Business use	Important caution
Bernoulli	One yes/no outcome	p	Conversion, default, defect indicator	One trial only; probability may vary by case.
Binomial	Number of successes in n trials	n, p	Conversions, defect counts	Requires independence and common p under basic model.
Poisson	Count per exposure interval	λ	Arrivals, incidents, failures	Basic model assumes mean equals variance.
Normal	Continuous measurement	μ, σ	Measurement models; sampling approximations	Do not infer normality from convenience.
t	Standardized sample-mean differences	degrees of freedom	Mean inference with estimated variance	Tail thickness depends on df.
Chi-square	Non-negative quadratic form	degrees of freedom	Categorical tests, variance-related inference	Approximation can be poor with small expected counts.

Distribution	Typical variable	Parameters	Business use	Important caution
F	Ratio of scaled variances	df1, df2	ANOVA and model comparison	Sensitive to assumptions in small samples.

Source/Note: Author's synthesis. Distribution choice follows the data-generating model and inferential target.

4.12 Sampling Distributions, Standard Errors and the Central Limit Theorem

A sample statistic is itself a random variable before the sample is observed. If repeated samples of the same size were drawn under the same design, the statistic would vary. Its sampling distribution describes that variation. This is the bridge from a single dataset to statistical uncertainty. The standard error is the standard deviation of the sampling distribution of an estimator.

$$SE(\bar{x}) = \sigma/\sqrt{n} \quad \text{or, when } \sigma \text{ is unknown, approximately } s/\sqrt{n}$$

Precision improves at the rate 1/√n, so gains from larger independent samples exhibit diminishing returns.

The central limit theorem (CLT) states, under broad conditions, that standardized sums or means tend toward a normal distribution as sample size increases, even when the underlying observations are not normal. The theorem is powerful but often overstated. It does not make the raw data normal. It does not cure dependence, selection bias or infinite-variance pathologies. And 'n = 30 is always enough' is a rule of thumb with no universal validity. Required sample size depends on skewness, tail behavior, dependence and the statistic being estimated.

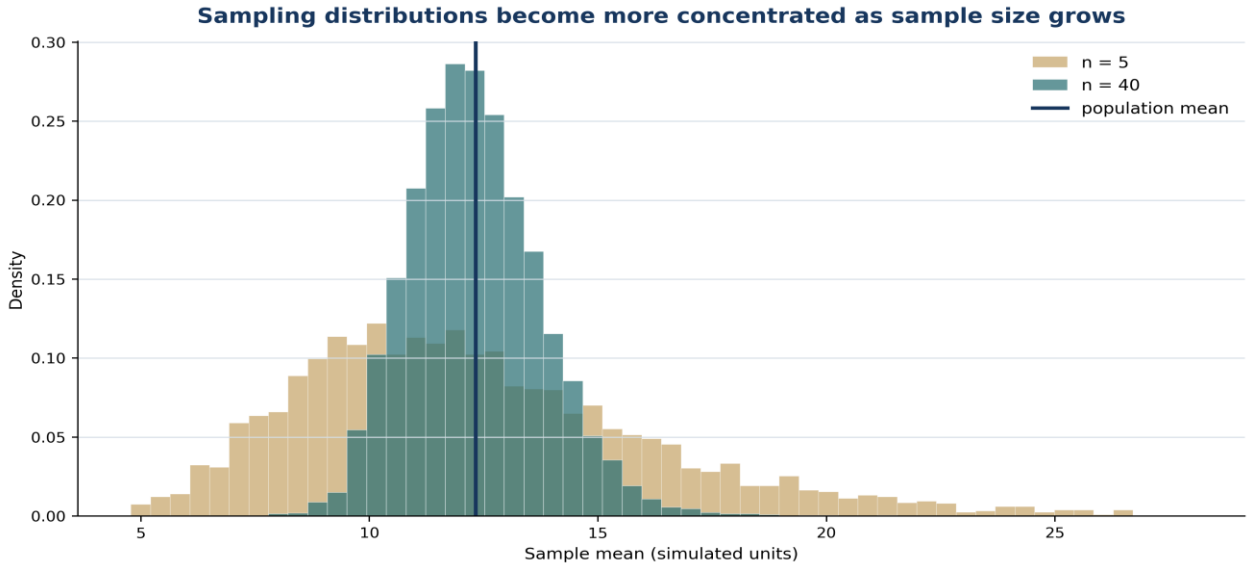

Figure 4.4. Simulated sampling distributions of the mean from a right-skewed population.

Source: Author's simulation. The underlying population is lognormal; 5,000 sample means were generated for each sample size.

Figure 4.4 shows two ideas. First, the distribution of sample means is more symmetric than the strongly skewed population. Second, the means from n = 40 vary less than those from n = 5. The population itself has not become less variable; the estimator has become more precise. This distinction between standard deviation and standard error is essential. Standard deviation describes spread among observations; standard error describes uncertainty in an estimated quantity.

4.12.1 Finite populations, clustering and effective sample size

The usual standard-error formulas assume a sampling design. When a large fraction of a finite population is sampled without replacement, a finite-population correction can reduce the standard error. When observations are clustered, however, effective information can be much smaller than the row count because units within a cluster resemble one another. A dataset with 10,000 transactions from 10 branches may contain far less independent information about branch-level effects than 'n = 10,000' suggests. Chapter 3 addresses sample design; Chapter 4 emphasizes that inferential formulas must match it.

4.13 Estimation and Confidence Intervals

4.13.1 Point estimates and interval estimates

A point estimate gives a single best estimate under a method, such as a sample mean for a population mean. An interval estimate expresses the range of parameter values compatible with the data and procedure at a chosen confidence level. Confidence intervals are often more informative than significance tests because they show magnitude and precision together.

$$\text{Estimate} \pm \text{critical value} \times \text{standard error}$$

For a mean with unknown variance under standard assumptions, use a t critical value; interval methods for proportions should respect boundary and sample-size conditions.

4.13.2 Correct interpretation of a 95% confidence interval

Under the frequentist interpretation, a 95% confidence procedure is designed so that 95% of intervals produced by repeated applications under the model would contain the fixed true parameter. Once a particular interval is computed, the parameter is not treated as a random draw from that interval. Saying ‘there is a 95% probability the true mean lies between 48 and 57’ is therefore not the standard frequentist interpretation. In professional communication, a less technical but defensible phrasing is that the interval provides a range of parameter values consistent with the data and procedure at the 95% confidence level.

WORKED EXAMPLE | A MEAN AND ITS UNCERTAINTY

A delivery sample has n = 36, mean 53 minutes and sample SD 12 minutes. The standard error is 12/√36 = 2 minutes. With 35 degrees of freedom, a two-sided 95% t critical value is about 2.03. The 95% confidence interval is therefore 53 ± 2.03(2), or approximately 48.9 to 57.1 minutes. Reporting only ‘mean = 53’ hides the precision of the estimate.

4.13.3 Confidence level, sample size and practical precision

A higher confidence level produces a wider interval, all else equal. Larger samples narrow intervals approximately with 1/√n, so halving a standard error generally requires about four times as many independent observations. Managers should therefore frame sample size around decision-relevant precision rather than ritual targets. A ±0.2 percentage-point interval may be necessary for a high-volume pricing experiment while ±5 points may be adequate for exploratory market research. Precision should be linked to consequences.

4.13.4 Proportions, differences and precision for rates

Many business outcomes are proportions rather than means: on-time deliveries, conversions, defaults, defects or customers retained. If a simple random sample contains n independent binary observations with sample proportion $\hat{p}$, the standard error is approximately the square root of $\hat{p}(1 - \hat{p})/n$. This approximation is most useful away from the boundaries 0 and 1 and when the effective sample size is adequate. Weighted, clustered or repeated observations require design-aware standard errors rather than the simple formula.

$$SE(\hat{p}) \approx \sqrt{[\hat{p}(1 - \hat{p}) / n]}$$

For an illustrative large-sample interval, $\hat{p}$ ± critical value × $SE(\hat{p})$. Near 0 or 1, or with small samples, use a better binomial interval rather than a mechanical Wald interval.

For example, if 420 of 500 independently sampled deliveries meet a contractual on-time definition, $\hat{p}$ = .84 and the simple standard error is about .0164. A 95% normal-approximation interval is therefore roughly .808 to .872. The interval expresses sampling precision under the assumptions; it does not account for a biased definition of “on time,” duplicated shipments, route clustering or a non-representative period. Precision is valuable only after the estimand and design are credible.

4.14 Hypothesis Testing: What a p-Value Can and Cannot Say

4.14.1 Null and alternative hypotheses

A hypothesis test evaluates how compatible observed data are with a specified null model relative to a defined test statistic. The null hypothesis commonly represents no difference, no association or a specified parameter value, while

the alternative represents departures considered relevant by the test. The test begins with a model and a decision rule; it does not begin with a search for $p < .05$.

4.14.2 Test statistics and p-values

A p-value is the probability, under the null model and other test assumptions, of obtaining a test statistic at least as incompatible with the null as the one observed. It is not the probability that the null hypothesis is true, not the probability that the result occurred 'by chance,' and not the size or importance of an effect. The American Statistical Association's statement emphasized precisely these limitations and warned against basing scientific, business or policy conclusions only on whether a p-value crosses a threshold (Wasserstein & Lazar, 2016).

The convention $\alpha = .05$ is historically common but should not be treated as a law of nature. A threshold is a decision convention that should reflect error costs, prior evidence, multiplicity and the stakes of action. Wasserstein, Schirm, and Lazar (2019) argued for moving beyond bright-line thinking and toward richer interpretation of evidence, uncertainty and practical importance. The appropriate response is not to ban p-values but to stop asking them to answer questions they were not designed to answer.

4.14.3 Type I error, Type II error and power

A Type I error occurs when a testing procedure rejects a true null hypothesis. Its long-run probability is controlled by α under the test assumptions. A Type II error occurs when the procedure fails to reject a false null hypothesis for a specified effect. Power is the probability of correctly rejecting the null for that effect. Power increases with larger true effects, larger effective samples, lower noise and, for fixed sample size, a more permissive α. These are trade-offs, not moral labels: lowering α reduces false positives but can increase false negatives unless information increases.

$$\text{Power} = 1 - \beta$$

β is the Type II error probability for a specified alternative, design and analysis plan.

4.14.4 One-sided and two-sided tests

A two-sided test allows departures in either direction. A one-sided test concentrates rejection probability in one direction and can have greater power there, but it is defensible only when the opposite direction would not change the inferential claim and when direction is specified before seeing the data. Switching to a one-sided test after observing a favorable direction is a form of result-driven analysis.

4.14.5 Fisherian evidence, Neyman-Pearson error control and the modern hybrid

Modern null-hypothesis significance testing often combines ideas that arose from distinct statistical traditions. In Fisherian significance testing, the p-value functions as a graded measure of how surprising the data are under a null model and can motivate further scrutiny; it was not originally a mechanical accept-reject rule. Neyman and Pearson instead formalized decision procedures that specify competing hypotheses and control long-run error rates through quantities such as alpha, beta and power (Fisher, 1925; Neyman & Pearson, 1933).

The distinction matters in professional practice because a fixed threshold answers a different question from a graded evidential assessment. A pre-specified rule can be appropriate when repeated operational decisions require explicit error control, while exploratory or scientific interpretation usually requires effect estimates, uncertainty, model criticism and replication. Treating $p < .05$ simultaneously as proof of an effect, a measure of importance and a universal decision rule mixes these traditions and overstates what the statistic can establish (Wasserstein & Lazar, 2016).

EVIDENCE CHECK | NON-SIGNIFICANT DOES NOT MEAN "NO EFFECT"

A p-value above .05 can arise because the true effect is small, the sample is noisy, the sample is too small, the model is misspecified or the null is approximately compatible with the data. The correct question is whether the confidence interval excludes effects large enough to matter, not whether software prints "NS."

4.15 t-Tests for Means

4.15.1 One-sample t-test

The one-sample t-test compares a sample mean with a specified benchmark μ_0 when the population standard deviation is unknown. The test statistic scales the observed difference by its standard error. The method assumes

independent observations and, for exact small-sample inference, a normally distributed population. It is often reasonably robust to moderate non-normality when sample size is adequate and extreme observations are absent.

$$t = (\bar{x} - \mu_0) / (s/\sqrt{n})$$

For the classical one-sample t-test, degrees of freedom are n - 1.

WORKED EXAMPLE | SERVICE TIME AGAINST A STANDARD

Using the previous sample, $\bar{x} = 53$, $s = 12$, $n = 36$. Testing H_0: $\mu = 48$ gives $t = (53 - 48)/2 = 2.50$ with 35 degrees of freedom. The two-sided p-value is about .017. The result is statistically incompatible with a 48-minute mean under the model, but management should still ask whether an estimated five-minute difference is operationally important and whether the sample represents the service process of interest.

4.15.2 Independent-samples t-test and Welch's test

An independent-samples t-test compares means from two independent groups. The classical pooled-variance version assumes equal population variances. Welch's (1947) test relaxes that equality assumption and is generally a safer default when group variances or sample sizes differ. The substantive assumptions remain: groups must be meaningfully comparable for the claim, observations should be independent within the design, and extreme skew or influential outliers can undermine small-sample inference.

A statistically significant difference between branches does not prove that branch management caused the difference. Assignment, geography, customer mix and workload may confound the comparison. Chapter 3 determines whether a causal comparison is defensible. The t-test only quantifies a mean difference under the specified design and model.

4.15.3 Paired t-test

Paired data arise when the same unit is measured twice or when observations are deliberately matched. The paired t-test operates on within-pair differences, not on two independent means. Examples include employee productivity before and after training, the same stores under two layouts, or matched supplier pairs. Pairing can increase precision by removing stable between-unit variation, but only if the pairing is genuine. Treating unrelated observations as paired creates false dependence; ignoring true pairing discards information.

4.16 Chi-Square Tests for Categorical Data

4.16.1 Goodness of fit

A chi-square goodness-of-fit test compares observed category counts with expected counts under a specified distribution. The statistic accumulates squared deviations scaled by expected counts. Large deviations produce a larger statistic and smaller p-value under the null model.

$$\chi^2 = \sum (O - E)^2 / E$$

O is an observed count and E the expected count. Large-sample validity depends on expected counts and the sampling design.

4.16.2 Test of independence

A chi-square test of independence assesses whether two categorical variables are statistically associated in a contingency table. Expected counts are computed from row and column margins under independence. The test says whether the table is more incompatible with independence than expected from sampling variation; it does not reveal which cells drive the association, how large the association is, or whether one variable causes the other. Standardized residuals and an effect-size measure such as Cramér's V can supplement the test.

When expected cell counts are small, exact or alternative methods may be preferable to the asymptotic chi-square approximation. Analysts should also distinguish counts from weighted survey estimates. Applying a naive chi-square test directly to complex survey data can understate uncertainty because stratification, clustering and survey weights alter the sampling distribution.

4.17 Analysis of Variance: Comparing More Than Two Means

4.17.1 Why not run many t-tests?

When three or more group means are compared, repeatedly testing every pair inflates the chance of at least one false positive. One-way analysis of variance (ANOVA) tests a global null hypothesis that the population means are equal under the model. It partitions variation into between-group and within-group components and compares their mean squares through an F statistic, a framework developed systematically in Fisher's early statistical work (Fisher, 1925).

$$F = MS_{between} / MS_{within}$$

A large F means between-group variation is large relative to within-group variation under the null model.

4.17.2 Assumptions and robust alternatives

Classical one-way ANOVA assumes independent observations, normally distributed errors within groups and equal variances. The method is often robust to moderate normality departures in balanced samples, but unequal variances combined with unequal group sizes can distort inference. Welch ANOVA relaxes the equal-variance assumption. If the outcome is strongly ordinal, heavily skewed or dominated by outliers, rank-based or robust alternatives may be appropriate, although each answers a somewhat different question.

4.17.3 Post-hoc comparisons and substantive interpretation

A significant global ANOVA does not identify which groups differ. Post-hoc procedures such as Tukey's honestly significant difference control family-wise error for pairwise comparisons under their assumptions. Planned contrasts can be more efficient when specific comparisons were defined before analysis. As always, report group means, intervals and effect sizes rather than presenting the F test alone.

4.17.4 When classical assumptions are poor: robust, rank and permutation alternatives

Classical procedures are not the only defensible options. Rank-based methods such as Wilcoxon-type comparisons and the Kruskal-Wallis test can be useful for ordinal outcomes or strongly non-normal data, while permutation tests can exploit the randomization structure of an experiment and robust estimators can reduce sensitivity to extreme observations. These methods do not mean "assumption free." They target different parameters or rely on different exchangeability conditions. The analyst should state what quantity the alternative procedure estimates and why that quantity answers the business question better than forcing an ill-fitting classical model.

Resampling methods provide another route to uncertainty when analytic reference distributions are inconvenient or fragile. A permutation test constructs a null reference distribution by rearranging labels in a way justified by randomization or exchangeability, whereas the bootstrap repeatedly resamples observed units to approximate the sampling distribution of a statistic (Efron & Tibshirani, 1993). Neither method is assumption-free: resampling must preserve clustering, pairing, time order or survey structure when those features are part of the data-generating process. Resampling can quantify uncertainty around a biased sample just as precisely as an analytic formula, so design quality remains prior to computation.

Table 4.7. Foundational inferential methods and their questions

Method	Primary question	Outcome type	Key design/assumption issues	Useful companion output
One-sample t	Does a mean differ from a benchmark?	Continuous	Independent observations; distributional robustness	Mean difference, CI, effect size.
Paired t	Is the mean within-pair difference nonzero?	Continuous paired	Correct pairing; distribution of differences	Mean difference, CI.
Welch two-sample t	Do two independent means differ?	Continuous	Independent groups; outliers; comparability	Difference, CI, standardized effect.
Chi-square independence	Are two categorical variables associated?	Categorical	Expected counts; survey design	Cell residuals, Cramér's V.
One-way ANOVA	Do 3+ group means differ globally?	Continuous	Independence; variance structure; outliers	Group CIs, η^2/ω^2, post-hoc contrasts.

Source/Note: Author's synthesis. Method selection must also respect the research design established in Chapter 3.

4.18 Effect Size, Multiple Testing, Practical Significance and Statistical Power

4.18.1 Effect size and decision relevance

Statistical significance depends on effect magnitude, variability and sample size. With very large datasets, commercially trivial differences can produce extremely small p-values. Effect-size measures describe magnitude. For two means, a raw difference may be the most interpretable effect. Standardized differences such as Cohen's d can help compare across scales but can obscure practical units. For categorical associations, risk differences, risk ratios or Cramér's V can be useful. For ANOVA, η^2 or ω^2 summarize the proportion of variance associated with group membership under the model. Cohen (1988) provided widely used conventional benchmarks, but context-specific thresholds are superior to universal labels such as 'small' or 'large.'

4.18.2 Multiple comparisons and false discoveries

If 100 independent null hypotheses are tested at $\alpha = .05$, about five rejections are expected by chance on average, although the realized number varies. This is why large dashboards and experimentation systems face multiplicity problems. Family-wise procedures such as Bonferroni control the probability of any false rejection but can be conservative. Benjamini and Hochberg's (1995) false-discovery-rate procedure controls the expected proportion of false discoveries among rejected hypotheses under specified conditions, often providing greater power in large testing families.

The modern scale of business experimentation makes this problem concrete. Microsoft's Experimentation Platform noted in July 2026 that individual A/B tests can generate hundreds or thousands of metric comparisons, creating a substantial risk that apparently significant movements are chance findings if correlation and multiplicity are ignored (Qi & Meng, 2026). The methodological lesson is durable: the more questions an analyst allows the data to answer, the stronger the need to account for the search process.

4.18.3 Power is not post-hoc evidence quality

Power is most useful in design, where it helps determine whether a study has a reasonable probability of detecting a decision-relevant effect. After observing results, a low p-value already contains information about the data under the null; calculating 'observed power' from the same effect estimate adds little and can be misleading. Post-study interpretation should emphasize effect estimates, confidence intervals, design quality and sensitivity rather than retrospective power labels.

CRITICAL DEBATE | SHOULD P < .05 DISAPPEAR?

Critics of threshold-based inference argue that binary significance labels encourage publication bias, p-hacking and exaggeration. Defenders note that formal error control remains useful when decisions require pre-specified rules. A defensible professional approach is to retain testing where it serves a clearly defined decision process while refusing to let a threshold replace effect size, uncertainty, replication, design quality and substantive judgment.

4.19 Correlation, Association and the Boundary of Causal Claims

4.19.1 Covariance and Pearson correlation

Covariance indicates whether two quantitative variables tend to vary together, but its magnitude depends on the units. Pearson's correlation coefficient standardizes covariance to a scale from -1 to +1. Values near +1 indicate strong positive linear association, values near -1 strong negative linear association and values near 0 weak linear association. Correlation is dimensionless but is sensitive to outliers and can miss strong nonlinear relationships.

$$r = \operatorname{cov}(X, Y) / (s_x s_y)$$

Pearson r measures linear association. Inspect the scatterplot before treating r as a sufficient summary.

Correlation does not imply causation because confounding, reverse causation, selection and common trends can create association. Nor does 'correlation does not imply causation' mean correlations are useless. They are valuable descriptive evidence and can be components of causal analysis when combined with appropriate design and assumptions. Chapter 3 established the counterfactual logic of causal inference; Chapter 4 preserves that boundary.

4.19.2 Spearman rank correlation

Spearman's rank correlation measures monotonic association using ranks. It is useful when variables are ordinal, when relationships are monotonic but nonlinear, or when extreme values make Pearson correlation unstable. It does not solve all outlier or dependence problems and should not be selected merely because a normality test rejects. The analytical question and data structure should govern the choice.

EVIDENCE CHECK | TIME TRENDS CAN CREATE SPURIOUS CORRELATION

Two unrelated business series can correlate strongly because both trend upward over time. Correlating raw monthly sales with a steadily increasing time index can therefore produce a high r even when no causal mechanism links the variables. Time-series structure must be addressed before ordinary cross-sectional interpretation is applied.

4.20 Simple and Multiple Linear Regression

4.20.1 What linear regression estimates

Linear regression models the conditional mean of a quantitative outcome as a linear function of one or more predictors. In simple regression, the slope β1 represents the expected change in the outcome associated with a one-unit increase in X under the model. In multiple regression, each coefficient is interpreted conditionally on the other included predictors, provided the model is correctly specified for that interpretation.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

β_0 is the intercept, β_1 the slope, and ε_i the residual component not explained by the fitted conditional mean.

Ordinary least squares (OLS) chooses coefficients that minimize the sum of squared residuals. The fitted line therefore summarizes average association in a precise mathematical sense. It does not guarantee good prediction, correct causal interpretation or adequate fit in every region of the data. Residual diagnostics and substantive knowledge remain essential.

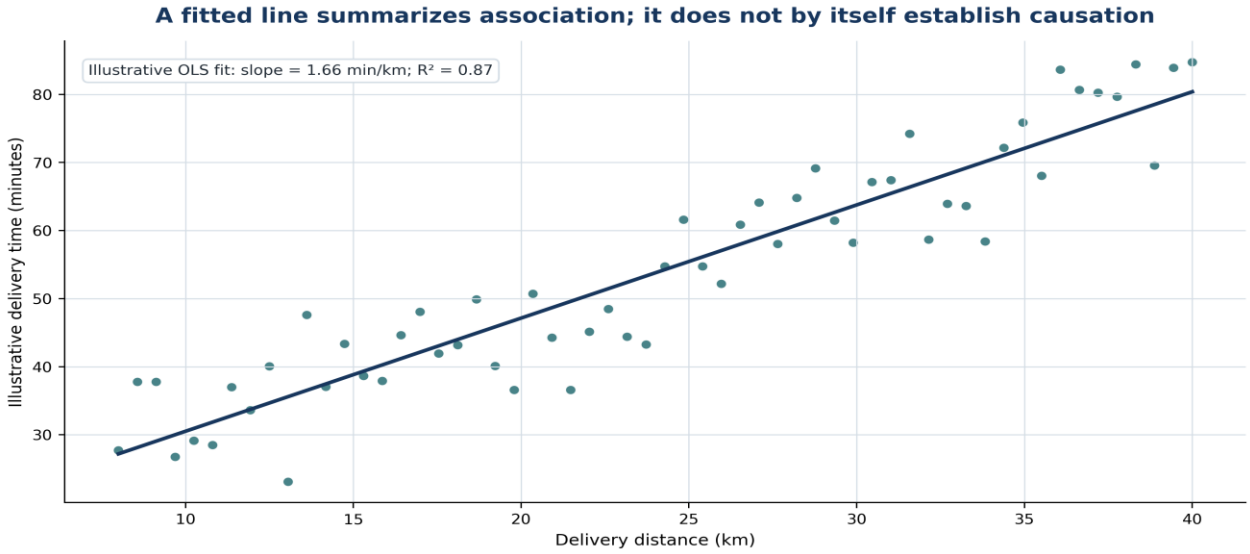

Figure 4.5. Illustrative linear association between delivery distance and delivery time.

Source: Author's synthetic data. The fitted line is descriptive and inferential only under stated assumptions; it does not establish a causal effect of distance.

4.20.2 Interpreting coefficients, R^2 and adjusted R^2

In the synthetic data underlying Figure 4.5, the fitted slope is approximately 1.66 minutes per kilometer and the intercept about 13.9 minutes, with R^2 about .87. The slope means that the fitted conditional mean increases by about 1.66 minutes for each additional kilometer within the observed range. The intercept represents the fitted value at zero kilometers, which may lie outside the meaningful data range and should not automatically receive substantive interpretation. R^2 is the proportion of sample variation in Y explained by the fitted linear model relative to a mean-only benchmark. A high R^2 can coexist with biased coefficients, poor extrapolation or omitted confounding.

$$R^2 = 1 - SSE/SST$$

SSE is the sum of squared residuals; SST is total variation around the sample mean. Adjusted R^2 penalizes additional predictors relative to sample size.

4.20.3 Multiple regression and conditional interpretation

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i$$

Each slope is a partial association conditional on the other included predictors. Causal interpretation requires stronger design and identification assumptions.

Multiple regression can adjust for measured differences, improve precision and estimate conditional associations. Yet 'controlling for variables' is not automatically beneficial. Conditioning on variables caused by both exposure and outcome, or on mediators when estimating total effects, can introduce bias. Selecting controls solely because they are statistically significant is also poor practice. Variable inclusion should follow the research question, theory and causal structure established before modeling.

4.20.4 Core regression assumptions

For unbiased OLS coefficient estimates under the classical model, the conditional mean must be correctly specified and the error must have zero conditional expectation given the predictors. Independence assumptions depend on the design. Constant variance, or homoskedasticity, is required for the usual OLS standard errors, not for the coefficient estimates themselves; heteroskedasticity-consistent standard errors can provide more reliable inference under changing variance (White, 1980). Normal residuals are not required for OLS coefficients and become mainly relevant to exact small-sample inference under classical assumptions. Multicollinearity inflates uncertainty when predictors carry overlapping information but does not by itself bias coefficients.

Table 4.8. Regression diagnostics and what they actually diagnose

Issue	Diagnostic signal	Consequence	Possible response
Nonlinearity	Curved residual pattern; structured scatterplot	Misleading average slope	Transform, add justified nonlinear terms, restrict range.
Heteroskedasticity	Residual spread changes with fitted values	Usual standard errors can be wrong	Use robust SE; reconsider variance/model scale.
Influential observations	Large leverage and residual influence	Coefficients depend strongly on few cases	Verify data; report sensitivity; use robust methods if justified.
Multicollinearity	Large SEs; unstable coefficients; high VIF	Imprecise conditional effects	Reframe variables, combine measures, collect more informative data.
Dependence	Residuals clustered by unit/time	SEs typically understated	Use cluster/time-aware methods appropriate to design.
Omitted confounding	Substantive causal pathway absent	Biased causal interpretation	Improve design/model; do not label association causal.

Source/Note: Author's synthesis. Advanced causal modeling and machine-learning extensions are beyond this chapter's boundary.

4.20.5 Categorical predictors, interactions and extrapolation

Categorical predictors are represented through indicator variables relative to a reference category. The coefficient is a conditional mean difference from that reference under the model. Interaction terms allow the association of one predictor to depend on another. An interaction is often substantively important even when neither main effect is interpretable in isolation at arbitrary zero values. Centering variables can improve interpretability but does not change the underlying fit of a linear model with the same terms.

Extrapolation is a recurring professional risk. A regression estimated among firms with 10 to 200 employees cannot safely predict a 5,000-employee enterprise merely because the formula accepts the number. Statistical models learn relationships over observed support. Crossing institutional regimes, technologies or policy periods can make historical coefficients irrelevant even within the numerical range.

MANAGERIAL INSIGHT | REGRESSION COEFFICIENTS ARE CONDITIONAL STATEMENTS, NOT UNIVERSAL LAWS
Before using a coefficient, state: conditional on which variables, for which population, during which period, over what data range, under which model, and for what type of claim. If those conditions cannot be articulated, the coefficient is not decision-ready evidence.

4.21 Time-Series Foundations and Introductory Forecasting

4.21.1 Time order changes the statistical problem

Time-series data are observations indexed in time order. Unlike ordinary cross-sectional samples, adjacent observations are often dependent. Sales this month may resemble sales last month because customers, capacity and market conditions persist. This autocorrelation means that ordinary formulas assuming independent rows can understate uncertainty. Time also introduces structural features such as trend, seasonality, cycles, interventions and regime changes.

4.21.2 Level, trend, seasonality and irregular variation

Level describes the baseline magnitude of a series. Trend is a persistent long-run movement. Seasonality is a recurring pattern tied to a known calendar period, such as day-of-week or month-of-year. Cyclical variation occurs over less fixed horizons, often linked to broader economic conditions. Irregular variation contains shocks and unexplained noise. Decomposition is useful because different forecasting methods assume different ways these components combine.

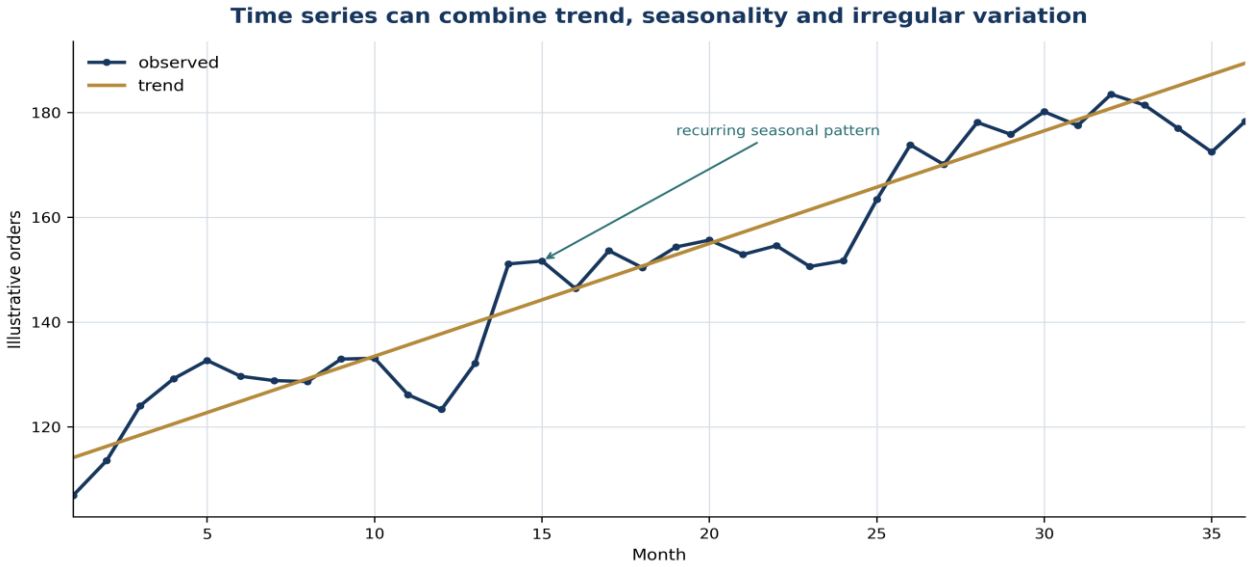

Figure 4.6. Illustrative monthly orders showing trend, seasonality and irregular variation.
Source: Author's synthetic data.

4.21.3 Benchmark forecasts

A forecast should be compared with a simple benchmark. The naïve method forecasts the next value as the most recent observed value. A seasonal naïve method forecasts each period using the corresponding value from the previous season. Moving averages smooth recent observations, while simple exponential smoothing gives declining weight to older observations. Trend methods can extend a local trend. These methods are intentionally introductory; Chapter 5 develops forecasting for managerial planning and more advanced predictive modeling.

$$\text{Naïve forecast: } \hat{y}(T+h|T) = yT$$

For a seasonal naïve method with season length m, use the most recent observed value from the same seasonal position.

4.21.4 Forecast accuracy and honest evaluation

Forecast accuracy must be evaluated on observations not used to fit the model whenever the goal is out-of-sample prediction. A holdout period, rolling-origin evaluation or time-series cross-validation respects temporal ordering. Randomly shuffling time-series rows into training and test sets can leak future information into the past. Common accuracy measures include mean absolute error (MAE), root mean squared error (RMSE) and mean absolute percentage error (MAPE). MAPE becomes unstable or undefined near zero and can bias comparisons across scales; scaled errors can be preferable in some settings.

$$\text{MAE} = (1/n)\sum|y_t - \hat{y}_t| \quad | \quad \text{RMSE} = \sqrt{[(1/n)\sum(y_t - \hat{y}_t)^2]}$$

RMSE penalizes large errors more heavily. No single accuracy metric captures all business costs.

Forecasting competitions reinforce the value of empirical benchmarking. The M4 Competition evaluated 61 methods across 100,000 time series and highlighted the strong performance of combinations and the importance of prediction intervals rather than point forecasts alone (Makridakis et al., 2020). Hyndman and Athanasopoulos (2021) similarly emphasize understanding time-series structure before building forecasts. In business practice, the best model is not the one with the most fashionable label but the one that performs credibly against relevant benchmarks under the intended forecast horizon and loss function.

4.21.5 Forecast intervals and structural breaks

A point forecast suppresses uncertainty. Forecast intervals widen as the horizon increases because more future shocks can accumulate and because parameter uncertainty matters. Historical intervals can also be misleading during structural breaks: pandemics, wars, regulatory reforms, currency crises, major competitor entry or technology shifts can change the process itself. Forecasting therefore requires both statistical methods and institutional awareness. Chapter 5 extends this foundation into scenario analysis and decision-making under risk.

4.22 African Case: Rwanda Enterprise Survey Data and the Discipline of Weighting

The World Bank Enterprise Surveys provide a useful bridge between Chapter 3's research-design logic and Chapter 4's statistical analysis. The Rwanda 2023-2024 survey dataset includes structured firm-level records collected under the Enterprise Surveys framework. One public metadata page for the variable asking firms to identify their biggest obstacle reports 358 valid records and shows, among the raw cases, 96 coded as access to finance, 49 as access to land, 42 as tax rates and smaller counts for other categories. Crucially, the same metadata page warns that these case counts cannot be interpreted as population summary statistics (World Bank, 2025).

That warning is not a technical footnote. Enterprise Surveys use sampling designs intended to represent a defined universe of formal firms, commonly stratified by firm size, sector and location. When some strata are sampled at different rates, a raw proportion gives every sampled establishment equal influence even though sampled establishments can represent different numbers of firms in the target population. Survey weights correct for the design and, where applicable, nonresponse or other adjustments. Analysts who ignore weights can turn accurate individual records into an inaccurate population claim.

$$\text{Weighted mean or proportion} = \sum(w_i y_i) / \sum w_i$$

Survey weights have design-specific meanings. Do not invent, rescale or normalize them without understanding the survey documentation and estimand.

WORKED EXAMPLE | RAW AND WEIGHTED PERCENTAGES CAN DIFFER

Imagine an illustrative survey with 50 large firms and 50 small firms. Thirty large firms and 10 small firms report finance as their biggest obstacle, so the raw sample proportion is 40%. If large firms were deliberately oversampled and each sampled large firm represents one population firm while each sampled small firm represents nine population firms, the weighted estimate is [30(1)+10(9)]/[50(1)+50(9)] = 120/500 = 24%. The numbers are illustrative, not an Enterprise Survey estimate; the point is that sampling design changes population inference.

Complex survey inference can require more than multiplying by weights. Stratification and clustering affect standard errors. Specialized survey procedures estimate uncertainty using the sample design rather than treating weighted records as an ordinary simple random sample (Lumley, 2010). The broader lesson applies to national labour

surveys, household surveys, trade surveys and customer research: weighting changes the estimand represented by the sample, while variance estimation must reflect how the sample was obtained.

AFRICAN CONTEXT | DATA SCARCITY AND DATA ABUNDANCE CAN COEXIST

African firms may face missing administrative series, informal-market undercoverage or inconsistent identifiers while simultaneously generating enormous volumes of mobile-money, platform and telecom data. The professional challenge is therefore not simply to obtain more data. It is to understand coverage, ownership, representativeness, privacy and the institutions that govern how data can legitimately be linked and interpreted.

4.23 Comparative Global Case: Large-Scale Online Experimentation and the False-Positive Problem

Digital firms can randomize millions of users and measure outcomes almost instantly, creating an environment in which statistical power can be enormous. Controlled online experiments have therefore become a major application of business statistics. Microsoft has documented an experimentation platform used to evaluate web and software changes through randomized A/B tests, emphasizing both the scientific value of randomization and the technical and cultural requirements for trustworthy analysis (Kohavi et al., 2009; Kohavi et al., 2020).

Scale does not make inference automatic. In 2026 Microsoft researchers highlighted the challenge created when a single experiment produces hundreds or thousands of metric comparisons. With many correlated metrics, some statistically significant movements will arise by chance even if the feature has little meaningful effect. Analysts must define primary metrics, use guardrails, understand sample-ratio anomalies, address multiple testing, and interpret effect sizes rather than celebrating every p-value below .05 (Qi & Meng, 2026). This is a global version of the same lesson encountered in small datasets: the number of possible analyses is itself part of the evidence-generating process.

Online experiments also sharpen ethical questions. Users may not know that minor interface variants are being tested. The risk can be trivial in some product experiments and substantial in others, especially where financial, health, safety, political or vulnerable-user outcomes are involved. Kohavi et al. (2020) therefore treat ethics as a distinct concern in experimentation. Statistical validity does not determine moral legitimacy. A perfectly randomized experiment can still be unethical if exposure creates disproportionate risk, informed expectations are violated or the experiment exploits users who cannot reasonably avoid the intervention.

EVIDENCE CHECK | BIG N DOES NOT CURE BAD MEASUREMENT

Millions of observations can make a biased estimate extraordinarily precise. If a platform measures clicks when the real outcome is customer welfare, or excludes users who abandon before an event is logged, narrower confidence intervals merely increase confidence about the wrong quantity.

4.24 Faith, Ethics and Moral Reasoning: Statistical Honesty as Truthful Stewardship

Statistics is often presented as ethically neutral because formulas do not possess intentions. Yet statistical practice is performed by persons who decide what to count, which records to exclude, what uncertainty to display, which subgroup to highlight, which model to privilege and how strongly to state a conclusion. Those choices distribute attention, credibility and sometimes resources. Statistical ethics therefore concerns both technical validity and the moral use of evidence.

4.24.1 Truthfulness and the temptation to manufacture certainty

A Christian approach begins with truthfulness. The command against false witness in Exodus 20:16 extends beyond direct verbal lies to professional practices that create a false impression. Proverbs 11:1 condemns dishonest scales, an image with immediate relevance to measurement. An analyst who changes a denominator, hides a confidence interval, removes inconvenient observations without disclosure or presents a selected subgroup as the whole is using numerical form to distort reality. The fact that every individual calculation is correct does not make the overall communication truthful.

Theological claims should not be confused with statistical premises. Scripture does not determine whether Welch's t-test or a rank-based test has better frequentist properties under heteroskedasticity. Those are empirical and mathematical questions. Christian moral reasoning enters at the level of purpose, honesty, stewardship, respect for

persons and accountability: the analyst must seek the most truthful representation available and must not exploit technical asymmetry to mislead people who cannot audit the method themselves.

4.24.2 Humility about uncertainty

Statistical uncertainty is not an embarrassment to hide. Proverbs 18:13 warns against answering before listening, and the principle translates professionally into resistance to premature certainty. Confidence intervals, sensitivity analyses and alternative explanations are practices of epistemic humility when they accurately represent what is not known. Humility is not indecision; it is proportioning confidence to evidence.

A serious alternative ethical perspective reaches similar duties through different foundations. Deontological ethics can emphasize duties not to deceive; consequentialist reasoning can emphasize the harms created by misleading statistics; virtue ethics can emphasize honesty, prudence and intellectual courage; rights-based reasoning can emphasize privacy and non-discrimination. These traditions differ in ultimate grounding and in how they resolve conflicts, but they converge on the professional danger of treating people as objects to be manipulated through selective information.

4.24.3 Human dignity, privacy and disaggregated data

Human data are not merely rows. A customer dataset can contain financial vulnerability, location, family circumstances or behavioral patterns. Disaggregation can reveal hidden discrimination and improve fairness, yet it can also re-identify small groups. Privacy and equity can therefore conflict. The analyst should ask whether the variable is necessary, whether access is proportionate, whether results can be aggregated safely, whether the affected persons can reasonably anticipate the use, and whether a subgroup analysis could stigmatize a community through weak evidence.

4.24.4 Stewardship, reproducibility and accountability

Stewardship concerns what is entrusted and how faithfully it is handled. 1 Corinthians 4:2 links entrusted responsibility with faithfulness. In quantitative work, reproducible code, preserved raw data, documented transformations and clear assumptions are forms of accountability because they allow another competent person to inspect the chain from observation to claim. Reproducibility does not guarantee truth, but deliberate opacity weakens trust and makes error harder to correct.

4.24.5 Statistical evidence and moral judgment

A statistically estimated effect does not decide whether an action should be taken. Suppose a pricing experiment shows that personalized urgency messages increase conversion. The result is empirical. Whether the firm should deploy the messages depends on how they work: truthful reminders may be legitimate, while fabricated scarcity or targeting of vulnerable customers can be manipulative. Similarly, a predictive model may identify employees likely to resign, but secret surveillance or punitive use of the score can violate dignity even if prediction accuracy is high. Statistical evidence informs moral judgment; it does not replace it.

FAITH AND VOCATION | THE ANALYST'S VOCATION IS NOT TO MAKE THE NUMBERS WIN

The faithful analyst serves truth and neighbour by making reality clearer, including when the result disappoints a sponsor, weakens a preferred strategy or shows that the available evidence is insufficient. Professional courage can require saying, 'We do not know yet,' or 'This effect is statistically detectable but too small to justify the harm or cost.'

4.25 Professional Toolkit: The Statistical Evidence Protocol

A professional can use the following protocol before releasing a quantitative claim. It is intentionally cumulative. Failure at an early stage cannot be repaired by greater complexity later.

Table 4.9. Statistical Evidence Protocol for business and trade analysis

Stage	Question to answer	Minimum evidence product	Stop condition
1. Claim	What exactly are we trying to learn?	Estimand, population, unit, period, comparison	Question is vague or changes after seeing results without disclosure.
2. Provenance	Where did each field come from?	Source inventory and data lineage	Origin, coverage or ownership cannot be established.

Stage	Question to answer	Minimum evidence product	Stop condition
3. Governance	May we lawfully and ethically use these data?	Purpose, access, privacy and retention record	Use exceeds permission, necessity or acceptable risk.
4. Quality	Are data fit for this claim?	Profile of missingness, duplicates, ranges, consistency	Critical fields are unreliable or systematically absent.
5. Description	What does the distribution actually look like?	Tables, plots, center, spread, denominators	Anomalies or aggregation effects remain unexplained.
6. Method	Which statistical method matches the question and design?	Assumptions, formula/model, planned contrasts	Method assumes a design or scale not present in the data.
7. Uncertainty	How precise is the estimate?	SE, CI, p-value where relevant, sensitivity	Only point estimates or threshold labels are available.
8. Magnitude	Is the effect large enough to matter?	Raw effect and context-relevant effect size	Result is statistically detectable but operationally negligible.
9. Robustness	Would reasonable alternatives change the conclusion?	Diagnostics, sensitivity, subgroup/multiplicity review	Conclusion depends on one fragile specification.
10. Communication	Can a non-technical decision maker understand limits?	Decision-ready visual and plain-language interpretation	Presentation hides denominator, uncertainty or material caveats.
11. Reproducibility	Can another analyst reconstruct the result?	Code/query, versioned data, dictionary, output log	Result depends on undocumented manual steps.
12. Moral review	Should the proposed use be pursued?	Documented privacy, fairness, dignity and foreseeable-harm judgment	Technically valid action violates a moral or ethical red line.

Source/Note: Author's synthesis. The protocol separates statistical validity from legal and moral legitimacy while requiring all three for responsible professional use.

4.25.1 A concise decision tree for method selection

Begin with the outcome. If the objective is to summarize one variable, use descriptive statistics and a distributional visual. If comparing a continuous mean with one benchmark, consider a one-sample t procedure; with two independent groups, Welch's t-test is often appropriate; with paired observations, analyze within-pair differences; with three or more independent groups, consider ANOVA or a robust alternative. If the question concerns association between categorical variables, use contingency-table methods. If it concerns linear association among quantitative variables, inspect scatterplots and consider correlation or regression. If observations are ordered in time, use time-series methods rather than treating rows as exchangeable. At each branch, design and assumptions can overrule the nominal method.

PRACTICE TOOL | FIVE SENTENCES FOR A DECISION-READY STATISTICAL RESULT

1. State the population, period and outcome. 2. State the estimate in natural units. 3. State the uncertainty interval. 4. State whether the evidence is compatible with the benchmark or comparison under the model. 5. State the most important limitation or alternative explanation. Add a recommendation only after those five sentences are clear.

4.26 Common Misconceptions, Analytical Errors and Professional Pitfalls

Table 4.10. Recurring statistical misconceptions and corrective professional practice

Misconception	Why it fails	Better professional practice
More data automatically mean better evidence.	Volume reduces sampling error only under appropriate sampling and measurement conditions. Selection bias, confounding, dependence and systematic measurement error can remain, and large samples can estimate those biases very precisely.	Audit provenance and design before celebrating n.

Misconception	Why it fails	Better professional practice
Cleaning means removing unusual observations.	Cleaning corrects or transparently represents known data problems. A genuine extreme can be the most important customer, fraud event, disruption or market shock in the dataset.	Verify provenance first; retain genuine extremes and report sensitivity.
Missing values can safely be replaced by the mean.	Mean imputation treats uncertainty as if it did not exist, compresses variance and can weaken or distort relationships among variables.	Diagnose missingness, justify the method and test sensitivity.
The mean is the "correct" average.	Mean, median, weighted mean, geometric mean and percentiles answer different questions. Distribution shape, scale and decision purpose determine which summary is meaningful.	Choose the summary that matches the estimand and distribution.
A normal-looking histogram proves normality.	Graphical appearance is sample-dependent, and many procedures concern the distribution of errors or estimators rather than the raw observations themselves.	Inspect diagnostics relevant to the model, not a ritual normality label.
The central limit theorem fixes biased samples.	The CLT describes the sampling behavior of certain statistics under conditions. It cannot create representativeness, correct a bad measure or make dependent observations independent.	Separate precision from bias.
A 95% confidence interval gives a 95% probability that the fixed parameter lies inside this realized interval.	In frequentist inference, 95% refers to the long-run coverage of the procedure under its assumptions, not a probability assigned to a fixed parameter after the interval is observed.	Communicate coverage and assumptions accurately.
$p < .05$ proves the alternative hypothesis.	A p-value measures the compatibility of observed data with a specified null model. It does not report the probability that either hypothesis is true and it does not measure effect magnitude.	Report the estimate, interval, design and substantive meaning.
$p > .05$ proves there is no effect.	Non-significance can reflect low precision, high noise, model mismatch or a genuinely small effect. A wide interval can include both important benefit and important harm.	Ask which decision-relevant effects the interval excludes.
Statistical significance means business significance.	With large samples, tiny effects can be highly detectable while remaining commercially negligible, ethically unacceptable or too costly to implement.	Evaluate magnitude, costs, benefits, feasibility and consequences.
Correlation near zero means no relationship.	Pearson correlation summarizes linear association. Strong nonlinear, segmented or interaction patterns can exist even when r is close to zero.	Plot the data and inspect plausible functional form.
Regression controls for confounding if enough variables are added.	Adding variables mechanically can leave confounding untouched, introduce collider or overcontrol bias, inflate variance and destroy interpretability.	Select controls from the research question and causal structure.
High R^2 means the model is true.	R^2 is an in-sample fit measure relative to a mean-only benchmark. It says nothing by itself about causal identification, external validity, stability or extrapolation.	Use diagnostics, validation and design-based reasoning.
Forecast accuracy can be judged on the same data used to fit the model.	In-sample fit rewards flexibility and can conceal overfitting. Forecast credibility requires evaluation on genuinely future observations using a horizon and loss measure aligned with the decision.	Use holdout or rolling-origin evaluation and compare benchmarks.
A dashboard is objective because it is automated.	Definitions, filters, denominators, thresholds, colors and aggregation rules are human choices. Automation can reproduce bias consistently rather than remove it.	Govern metric definitions and make adverse results visible.
AI-generated statistical code is trustworthy if it runs.	Executable code can select the wrong estimand, mishandle weights, leak future data, violate privacy, use an invalid test or produce a confident but false interpretation.	Require human verification, tests, provenance and reproducibility.

Source/Note: Author's synthesis. The table identifies common interpretive failures and the professional discipline required to correct them.

4.27 Chapter Synthesis: From Designed Evidence to Statistical Evidence

Chapter 3 ended with evidence whose provenance, sampling logic and measurement purpose were understood. Chapter 4 has shown how quantitative observations become statistical evidence. The process begins with an estimand and data lineage, continues through governance, quality checks and reproducible cleaning, and only then reaches descriptive statistics, probability and inference. This order is methodological rather than cosmetic: bad data quality can invalidate a beautifully estimated model, while a well-defined dataset can often answer an important question with simple statistics.

Descriptive statistics compress distributions, but compression must preserve relevant structure. Means, medians, proportions, standard deviations, quantiles and graphs answer different questions. Exploratory analysis reveals patterns and anomalies but should be distinguished from pre-specified confirmatory inference. Visualization is part of statistical reasoning because axes, denominators and encodings can reveal or conceal evidence.

Probability explains uncertainty; sampling distributions explain why estimates vary; standard errors and confidence intervals quantify precision. Hypothesis tests compare observed evidence with null models but p-values do not measure truth, importance or causal effect. t-tests, chi-square tests and ANOVA provide structured comparisons when their design and assumptions fit. Effect size, multiplicity and power ensure that threshold-based testing does not substitute for substantive judgment.

Correlation describes association; regression estimates conditional relationships; neither becomes causal without a credible design and assumptions. Time series require respect for temporal order, dependence, trend and seasonality, while forecasts should be benchmarked out of sample and accompanied by uncertainty. Across all methods, reproducibility and ethical communication are part of technical quality rather than optional extras.

The Christian moral lens deepens these statistical duties by treating truth, stewardship, human dignity and accountability as intrinsic obligations. Numbers must not be used to manufacture certainty, hide inconvenient outcomes, invade privacy without justification or provide a technical veneer for manipulation. Alternative ethical traditions may ground these duties differently, but the professional implication is shared: analytical power increases responsibility.

4.27.1 Hand-off to Chapter 5

Chapter 4 has taught the reader how to describe data, quantify uncertainty and estimate statistical relationships. Chapter 5, Business Analytics and Decision Intelligence: Modeling, Prediction, Risk and Managerial Judgment, asks the next question: how should an organization choose and defend action under uncertainty? It will extend the statistical foundation into predictive models, classification, clustering, scenario analysis, optimization, simulation, structured decision methods, AI-supported judgment and feedback loops. The boundary remains deliberate. A statistically significant relationship does not tell a manager what objective to optimize, what risks to accept, what constraints are moral, or how to choose among competing values. Chapter 5 begins where statistical evidence ends: with accountable choice.

Key Terms

Table 4.11. Key terms and definitions

Term	Definition
Adjusted R^2	A version of R^2 that penalizes adding predictors relative to sample size and model complexity.
ANOVA	Analysis of variance; a family of methods comparing group means through between- and within-group variation.
Bayes' theorem	A probability rule for updating the probability of a condition using evidence and base rates.
Bias	Systematic error that moves an estimate away from the target quantity.
Central limit theorem	A set of results under which standardized sums or means approach a normal distribution under broad conditions as sample size increases.
Chi-square test	A test based on discrepancies between observed and expected categorical counts.
Confidence interval	An interval produced by a procedure with specified long-run coverage under its assumptions.

Term	Definition
Correlation	A standardized measure of association; Pearson correlation measures linear association.
Data governance	Rules, roles and controls governing data definition, quality, access, use, security, retention and accountability.
Data lineage	Traceable record of the source and transformations of analytical data.
Data profiling	Systematic inspection of types, ranges, missingness, duplicates, distributions and integrity before analysis.
Effect size	A quantitative measure of the magnitude of a difference, association or treatment effect.
Estimand	The precisely defined population quantity a statistical analysis seeks to estimate.
Exploratory data analysis	Numerical and graphical investigation of data structure, anomalies and plausible relationships.
Heteroskedasticity	Non-constant variance of regression errors across levels of predictors or fitted values.
Hypothesis test	A procedure evaluating how compatible observed data are with a specified null model using a test statistic and reference distribution.
Interquartile range	Q3 - Q1; the spread of the middle 50% of observations.
Missing at random	A missingness condition under which missingness can be explained by observed information under the model.
Multiple testing	Conducting a family of statistical tests, creating increased risk of false discoveries unless the search process is addressed.
Outlier	An observation unusually distant or influential relative to other observations under a defined criterion.
p-value	Under a specified null model, the probability of a test statistic at least as incompatible with the null as the observed statistic.
Practical significance	The substantive importance of an estimated effect for a real decision, distinct from statistical detectability.
Probability distribution	A formal description of probabilities assigned to possible values of a random variable.
Regression	A model relating the conditional mean of an outcome to one or more predictors.
Reproducibility	Ability to regenerate analytical results from documented data, code, assumptions and workflow.
Sampling distribution	The probability distribution of a statistic across repeated samples under a specified design.
Standard deviation	A measure of dispersion among observations in the original unit of measurement.
Standard error	The standard deviation of the sampling distribution of an estimator.
Statistical power	Probability that a testing procedure rejects the null for a specified alternative effect.
Time series	Observations indexed in chronological order, often exhibiting dependence, trend and seasonality.
Type I error	Rejecting a true null hypothesis under a testing procedure.
Type II error	Failing to reject a false null hypothesis for a specified alternative.
Weighted mean	An average in which observations contribute according to specified weights with a documented interpretation.

Source/Note: Definitions are specific to the usage in this chapter.

Concept-Check Questions

1. Why can a dataset be accurate but still be poor quality for a particular decision?
2. What is the difference between a unit of observation and a unit of analysis?
3. Why should raw data normally be preserved separately from processed data?
4. How do mean and median respond differently to extreme values?
5. What is the difference between standard deviation and standard error?
6. What does the central limit theorem not guarantee?
7. Give a correct interpretation of a 95% frequentist confidence procedure.
8. Why is a p-value not the probability that the null hypothesis is true?
9. When is Welch's independent-samples t-test preferable to the pooled-variance t-test?
10. What does a significant chi-square test of independence establish, and what does it not establish?
11. Why does running many hypothesis tests increase false-positive risk?
12. How can a regression have a high R^2 and still be unsuitable for decision-making?
13. Why should time-series observations not usually be randomly shuffled before forecast evaluation?
14. Why can a very large sample produce a statistically significant but commercially trivial result?
15. How do data privacy and reproducibility sometimes pull in different directions, and how can the conflict be managed?

Higher-Order Analytical and Critical-Thinking Questions

1. A national statistics office reports headline inflation of 12%, while your firm's procurement costs rose 25%. Explain why both figures can be correct and design an analysis to reconcile them.
2. An executive asks analysts to remove "abnormal" customers because their purchases make the average unstable. Develop a statistically and ethically defensible response.
3. A survey of 2,000 social-media followers finds 70% support for a new product. Under what conditions could this estimate be far less credible than a random sample of 400 customers?
4. Explain how a dashboard can be factually accurate yet systematically misleading without containing any false number.
5. Compare the managerial consequences of Type I and Type II errors in fraud detection, safety monitoring and marketing experimentation. Should α be the same in all three contexts?
6. A regression finds that firms using AI tools have 18% higher productivity after controlling for size and sector. Identify at least four alternative explanations that prevent immediate causal interpretation.
7. Why might a policy analyst prefer confidence intervals and effect sizes to a binary significance label when advising government on SME support?
8. Design a principled multiple-testing strategy for a platform experiment with one primary outcome, five guardrail metrics and 200 exploratory metrics.
9. Explain why a forecasting model that was best from 2018-2023 might fail in 2026 even if its historical RMSE was low.
10. Using Christian moral reasoning and at least one alternative ethical framework, evaluate whether a firm should use a statistically accurate model that predicts financial distress from customers' location and mobile-phone behavior.

Applied Case: LakeLink Distribution Ltd. - From Messy Operations Data to a Defensible Service-Level Claim

LakeLink Distribution Ltd. is a fictional East African distributor serving retailers across three corridors. Management believes delivery reliability deteriorated after fuel-price and route changes, but the operating dashboard shows only a single monthly on-time percentage. An analyst extracts 8,420 shipment records for the last two quarters. Initial profiling finds 164 duplicate shipment IDs, 3.8% missing promised-delivery timestamps, 41 negative recorded transit times caused by reversed timestamp fields, and 73 shipments with transit times above 10 days. The longest delays are concentrated in one border route rather than being random data-entry errors.

The dashboard defines 'on time' as delivery before midnight on the promised calendar date. However, contracts define on time as delivery within two hours of the promised timestamp. The dashboard also gives each shipment equal

weight even though some customers place consolidated high-value shipments and others place many small orders. Finally, the two quarters differ in route mix: the later quarter contains more shipments on the historically slower border corridor.

A statistically responsible analysis must therefore resolve the estimand before testing whether performance deteriorated. If the question concerns the share of individual shipments meeting contractual service, each shipment may be the appropriate unit, subject to repeated-customer and route dependence. If the question concerns customer experience, customer-level aggregation may be more relevant. If the question concerns revenue at risk, value weighting answers a different estimand. None is 'the' on-time rate independent of purpose.

Table 4.12. LakeLink illustrative summary after documented cleaning

Quarter	Shipments	Raw on-time %	Route-adjusted estimate	Mean transit hours	P90 transit hours
Q1	3,910	86.8%	85.9%	31.4	58.0
Q2	4,346	84.9%	85.4%	32.7	61.5

Source/Note: Fictional values for pedagogical analysis. "Route-adjusted estimate" is intentionally illustrative and does not specify a full model.

The raw on-time rate declines 1.9 percentage points, while a route-adjusted comparison suggests a smaller decline. Mean transit time rises 1.3 hours and the 90th percentile rises 3.5 hours, indicating that tail performance may have deteriorated more than the average. A single t-test on all shipment times would be inadequate because the outcome is skewed, routes differ, repeated customers create dependence and the contractual KPI is binary. The analyst should report descriptive distributions, define the comparison model, use design-appropriate standard errors and separate the evidence about change from the decision about adding capacity.

APPLIED CASE QUESTIONS

1. Which records should be corrected, excluded or retained, and why?
2. Define three different legitimate estimands for "delivery reliability."
3. Which visualizations would you use before formal inference, and what would each reveal?
4. What statistical method would you use for the contractual on-time outcome, and what dependence or design complications must be handled?
5. How would you examine whether deterioration is concentrated in the upper tail rather than the center of the transit-time distribution?
6. What additional evidence is required before attributing the change to fuel-price or route-policy changes?

Ethical and Faith-Informed Dilemma: The Executive Dashboard

A chief operating officer is preparing a board presentation after a difficult quarter. The official dashboard shows customer complaints up 28%, but the increase is statistically concentrated in two product lines affected by a known recall. The officer asks the analytics team to present complaints 'per active product' rather than 'per customer' because the revised denominator makes the increase only 6%. Both denominators can be calculated correctly. The board is not told that the metric definition changed. The officer argues that the revised metric better reflects the current portfolio and that alarming the board could jeopardize jobs and financing.

Evaluate the request. Which denominator best answers the board's likely question? When can a metric definition legitimately change? What disclosure is required? Does the possibility of protecting jobs justify a presentation that creates a materially different impression? Apply statistical communication principles, Christian duties of truthfulness and stewardship, and at least one serious alternative ethical framework. Your answer should distinguish a technically defensible redefinition from selective framing intended to conceal deterioration.

Data and Evidence Exercise: Rwanda CPI, January-July 2026

Table 4.13. Rwanda headline CPI inflation reported in 2026

Month	Year-on-year inflation (%)
January	8.9
February	9.2

Month	Year-on-year inflation (%)
March	9.2
April	13.0
May	12.9
June	13.6
July	14.5

Source/Note: National Institute of Statistics of Rwanda (2026a). These are year-on-year rates and should not be summed to obtain cumulative inflation.

Using Table 4.13, complete the following tasks. First, calculate the arithmetic mean, median, range and sample standard deviation of the seven reported year-on-year rates. Second, graph the series and describe the pattern without making a causal claim. Third, calculate the month-to-month change in the reported year-on-year rate, distinguishing a change in the inflation rate from a monthly price change. Fourth, estimate a simple linear trend across the seven observations and discuss why a forecast based on such a short series should be treated cautiously. Fifth, use the July 2026 detailed release to compare headline inflation with at least three component categories and explain why a firm-specific cost index could legitimately differ from headline CPI. Sixth, identify one ethical or communication risk that would arise if an analyst used only the component with the largest increase in a public presentation.

References

African Union. (2022). *AU data policy framework.* https://au.int/en/documents/20220728/au-data-policy-framework

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician, 27*(1), 17-21. https://doi.org/10.1080/00031305.1973.10478966

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289-300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Broman, K. W., & Woo, K. H. (2018). Data organization in spreadsheets. *The American Statistician, 72*(1), 2-10. https://doi.org/10.1080/00031305.2017.1375989

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association, 79*(387), 531-554. https://doi.org/10.1080/01621459.1984.10478080

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Efron, B., & Tibshirani, R. J. (1993). *An introduction to the bootstrap*. Springer. https://doi.org/10.1007/978-1-4899-4541-9

Fisher, R. A. (1925). *Statistical methods for research workers.* Oliver and Boyd.

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

Kohavi, R., Crook, T., Longbotham, R., Frasca, B., Henne, R., Lavista Ferres, J., & Melamed, T. (2009). *Online experimentation at Microsoft.* Microsoft Research. https://www.microsoft.com/en-us/research/publication/online-experimentation-at-microsoft/

Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to A/B testing.* Cambridge University Press. https://doi.org/10.1017/9781108653985

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260

Lumley, T. (2010). *Complex surveys: A guide to analysis using R*. Wiley. https://doi.org/10.1002/9780470580066

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2020). The M4 Competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting, 36*(1), 54-74. https://doi.org/10.1016/j.ijforecast.2019.04.014

National Institute of Statistics of Rwanda. (2026a). *Consumer Price Index (CPI) - 2026.* https://www.statistics.gov.rw/data-sources/surveys/Consumer-Price-Index/consumer-price-index-cpi-2026

National Institute of Statistics of Rwanda. (2026b, August 10). *Consumer Price Index (CPI) - July 2026.* https://www.statistics.gov.rw/statistical-publications/price-indices-cpi-ppi/consumer-price-index-cpi-july-2026

Neyman, J., & Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, 231*(694-706), 289-337. https://doi.org/10.1098/rsta.1933.0009

Qi, K., & Meng, M. (2026, July 15). *Treatment effect assessment at scale: Accounting for correlated metrics and metric relevance in modern experimentation.* Microsoft Research. https://www.microsoft.com/en-us/research/articles/treatment-effect-assessment-at-scale-accounting-for-correlated-metrics-and-metric-relevance-in-modern-experimentation/

Republic of Rwanda. (2021). Law No. 058/2021 of 13/10/2021 relating to the protection of personal data and privacy. Official Gazette, Special of 15/10/2021. https://dpo.gov.rw/fileadmin/DPO/Law_relating_to_the_protection_of_personal_data_and_privacy.pdf

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 581-592. https://doi.org/10.1093/biomet/63.3.581

Student. (1908). The probable error of a mean. *Biometrika, 6*(1), 1-25. https://doi.org/10.1093/biomet/6.1.1

Tukey, J. W. (1977). *Exploratory data analysis.* Addison-Wesley.

Wang, R. Y., & Strong, D. M. (1996). Beyond accuracy: What data quality means to data consumers. *Journal of Management Information Systems, 12*(4), 5-33. https://doi.org/10.1080/07421222.1996.11518099

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician, 70*(2), 129-133. https://doi.org/10.1080/00031305.2016.1154108

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05". *The American Statistician, 73*(sup1), 1-19. https://doi.org/10.1080/00031305.2019.1583913

Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika, 34*(1-2), 28-35. https://doi.org/10.1093/biomet/34.1-2.28

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica, 48*(4), 817-838. https://doi.org/10.2307/1912934

Wickham, H. (2014). Tidy data. *Journal of Statistical Software, 59*(10), 1-23. https://doi.org/10.18637/jss.v059.i10

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., Bourne, P. E., Bouwman, J., Brookes, A. J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C. T., Finkers, R., ... Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data, 3*, 160018. https://doi.org/10.1038/sdata.2016.18

World Bank. (2021). *World development report 2021: Data for better lives.* World Bank. https://doi.org/10.1596/978-1-4648-1600-0

World Bank. (2025). Rwanda - World Bank Enterprise Survey 2023-2024: Biggest obstacle affecting the operation of this establishment [Data catalog variable]. https://microdata.worldbank.org/index.php/catalog/6468/variable/V305

World Bank Enterprise Surveys. (n.d.). *Methodology*. Retrieved August 27, 2026, from https://www.enterprisesurveys.org/en/methodology

Annotated Further Reading

Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: Principles and practice* (3rd ed.). OTexts. https://otexts.com/fpp3/

A rigorous, accessible business-focused introduction to time-series exploration and forecasting. Particularly valuable for the Chapter 4-5 transition because it treats simple benchmarks, forecast evaluation and uncertainty carefully.

Kohavi, R., Tang, D., & Xu, Y. (2020). *Trustworthy online controlled experiments: A practical guide to A/B testing.* Cambridge University Press. https://doi.org/10.1017/9781108653985

A practice-rich treatment of experimentation at digital scale, including metrics, inference, organizational systems and ethical issues. Useful for understanding why statistical trustworthiness is a system property rather than a p-value.

Little, R. J. A., & Rubin, D. B. (2019). *Statistical analysis with missing data* (3rd ed.). Wiley. https://doi.org/10.1002/9781119482260

The definitive advanced reference on missing-data mechanisms and methods. Readers should use it to move beyond simplistic deletion and mean-imputation practices.

Tukey, J. W. (1977). *Exploratory data analysis.* Addison-Wesley.

A foundational work that shaped modern graphical and exploratory thinking. Its enduring lesson is to inspect data structures before forcing observations into a model.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p < 0.05". *The American Statistician, 73*(sup1), 1-19. https://doi.org/10.1080/00031305.2019.1583913

An essential entry point into contemporary debates about statistical significance, evidence, uncertainty and the misuse of bright-line thresholds.

World Bank. (2021). *World development report 2021: Data for better lives.* World Bank. https://doi.org/10.1596/978-1-4648-1600-0

A broad institutional treatment of data as an economic and governance resource. Especially useful for African and cross-border discussions of data value, rights, access, infrastructure and policy trade-offs.

END OF CHAPTER 4

The reader should now be able to turn well-designed quantitative data into transparent statistical evidence and to explain what that evidence does and does not justify. Chapter 5 converts this statistical foundation into structured decision intelligence under uncertainty.