1 *Statistics in Engineering*

This chapter explains what statistics is for and why an engineer needs it. The central fact is that engineering data vary: two trucks on the same route burn different amounts of fuel, two runs of the same perception model miss the same obstacle by different distances, and two visitors shown the same signup page make different choices. Statistics is the set of methods for collecting such data, describing them honestly, and drawing conclusions from a limited sample about the larger population the engineer actually cares about. The chapter is organized in the following order: the role of statistics in engineering; populations, samples, and the two directions of statistical reasoning; the three ways engineering data are collected; the numerical summaries of a data set (the sample mean, the sample variance, and the sample standard deviation); the graphical summaries (dot diagrams, histograms, and box plots); a roadmap of what this course covers; and a step-by-step procedure for statistical thinking that you will apply in every later chapter.

Pre-class quiz. *Try these before lecture; the answers follow at the end of the quiz.*

Q1. A quality engineer measures the diameter of 20 shafts taken from a production run of 5,000 shafts. Which set of shafts is the *population*, and which is the *sample*?

Q2. Two careful operators measure the hardness of the same weld ten times each and get slightly different numbers every time. Is someone making a mistake?

Q3. You are told that a coin is fair, and you compute the chance of seeing 8 or more heads in 10 tosses. Is this reasoning from population to sample, or from sample to population?

Answers.

Q1. The production run of 5,000 shafts is the population, because it is the whole collection the engineer wants a conclusion about. The 20 measured shafts are the sample: the subset actually observed.

Q2. Not necessarily. Repeated measurements of the same quantity almost always differ, because of small changes in the material, the instrument, and the operator. This spread is called variability, and it is normal. Statistics exists because variability exists.

Q3. From population to sample. The population behavior (a fair coin) is known, and you deduce what samples will look like. That is probability, the subject of ISE 205. This course runs the other direction: from an observed sample to statements about an unknown population. That is statistical inference.

Lecture 1 starts here — *Overview and course introduction*

Lecture 1. This section is covered by Lecture 1.

Reference. Montgomery & Runger, 6th ed., Sections 1-1 and 1-2.

Statistics. The science of collecting data, describing them, and drawing conclusions from a sample about the population it came from, in the presence of variability.

Variability. The fact that successive observations of a system or process do not produce exactly the same result.

1.1 *The Role of Statistics in Engineering*

Engineers solve problems by applying the engineering method: identify the problem, propose a model, run the model against data, and refine the design until it works. Every step of that loop that touches data touches statistics. When you ask whether a batch is within specification, how many defects to expect, whether a new process really improved yield, how long a component will last, or how many test cases are enough to trust an autonomous system, you are asking a statistical question.

The reason statistics is needed, rather than plain arithmetic, is variability. If every truck burned exactly the same fuel and every model run missed by exactly the same distance, one measurement would tell the whole story and no course would be required. In reality successive observations differ, and the differences carry information. Figure 1.1 shows five hourly output readings from each of two packing lines. Both lines average 10.0 tonnes per hour, yet Line B swings much more widely than Line A. A planner who schedules downstream work using only the average will be surprised by Line B far more often than by Line A. Please pay attention here, because this is a common trap: reporting an average without a

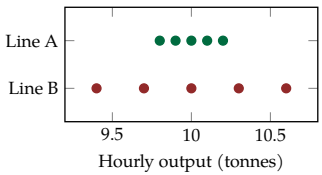

Figure 1.1. Two packing lines with the same average output but very different variability. The average alone cannot tell them apart.

measure of spread hides exactly the information an engineer needs for capacity, safety margins, and guarantees.

Insight. Variability is not noise to be ignored; it is a quantity to be measured. Almost every method in this course—confidence intervals, hypothesis tests, regression, designed experiments—is at heart a comparison between an effect the engineer hopes is real and the variability that could have produced the same appearance by chance. If you can measure the variability, you can tell the difference.

Engineering judgment. Variability has physical causes: tool wear, ambient temperature, raw-material lots, operator technique, sensor noise. Your engineering knowledge of the process tells you which causes are plausible; statistics tells you how large their combined effect is.

Example 1.1 — Two trucks, one route

A logistics company runs two trucks of the same model on the same 420 km route. Over five trips each, truck 1 used 31.9, 32.4, 31.6, 32.8, and 32.3 liters per 100 km, while truck 2 used 30.1, 34.0, 31.2, 33.5, and 32.2 liters per 100 km. A dispatcher looks at the averages, finds both close to 32.2, and concludes the trucks are ''the same.'' What is missing from that conclusion? Name at least three physical sources of the trip-to-trip differences.

Solution. Both trucks do average about 32.2 liters per 100 km, but truck 1's trips stay within about ± 0.6 of its average while truck 2's trips range over almost 4 liters per 100 km. The trucks have similar *centers* but very different *spreads*, so they are not the same for planning purposes: fuel budgets and range guarantees built on truck 2's average will fail more often. Plausible physical sources of the variability include traffic and average speed, load weight, wind and temperature, tire pressure, and driver behavior. The rest of this chapter builds the two numbers that make this comparison precise: the sample mean for the center and the sample standard deviation for the spread.

Reference. Montgomery & Runger, 6th ed., Sections 1-1 and 7-1.

1.2 *Populations, Samples, and Two Directions of Reasoning*

Every statistical problem in this course involves two collections. The *population* is the whole collection the engineering question is about: all 5,000 shafts in the run, all containers that will pass through the port this year, all future users who will ever see the signup page, all driving situations an autonomous vehicle will ever

Population. The entire collection of items, measurements, or outcomes that the engineer wants a conclusion about.

Sample. The subset of the population that is actually observed or measured.

encounter. The *sample* is the part we actually observe: 20 measured shafts, 40 tracked containers, 500 visitors in a pilot, 1,200 logged test scenarios. We study the sample because studying the population is impossible, too expensive, too slow, or destructive.

Parameter. A numerical property of the population, such as μ, σ^2, or p. Parameters are fixed but usually unknown.

Statistic (formal). A quantity computed from sample data, such as $\bar{x}$, s^2, or $\hat{p}$. Statistics are known once the data are in hand, but they change from sample to sample.

Each collection has its own numerical summaries, and the notation keeps them apart. A number that describes the population is a *parameter* and is written with a Greek letter: the population mean μ, the population variance σ^2, the population proportion p. A number computed from the sample is a *statistic* and is written with a Latin letter or an accented symbol: the sample mean $\bar{x}$, the sample variance s^2, the sample proportion $\hat{p}$. Table 1.1 lists the pairs you will use all semester. Please keep this distinction sharp from the first week: the parameter is the fixed, unknown target; the statistic is the computable, variable arrow we shoot at it.

Parameter	What it measures	Statistic that estimates it
μ	Mean of a population	$\bar{X}$
σ^2	Variance of a population	S^2
σ	Standard deviation of a population	S
p	Proportion of a population	$\hat{p}$
$\mu_1 - \mu_2$	Difference in two population means	$\bar{X}_1 - \bar{X}_2$
$p_1 - p_2$	Difference in two population proportions	$\hat{p}_1 - \hat{p}_2$

Table 1.1. Population parameters and the sample statistics that estimate them. Greek letters are unknown targets; the statistics are computed from data.

Statistical inference. Drawing conclusions about a population from the information in a sample, with a statement of how uncertain the conclusion is.

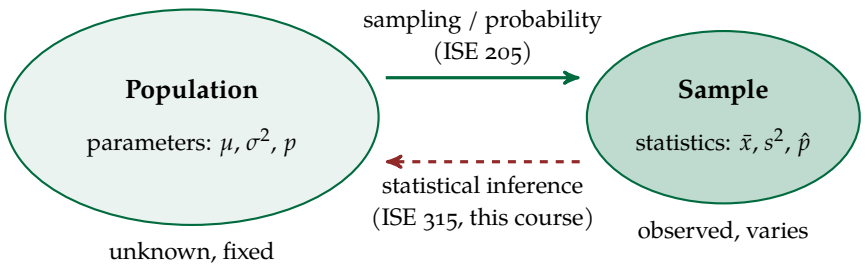

Figure 1.2. The two directions of statistical reasoning. Probability deduces sample behavior from a known population; inference reasons back from an observed sample to the unknown population.

Figure 1.2 shows the two directions of reasoning, and the distinction is the cleanest way to see where this course sits relative to its prerequisite. In ISE 205 the population was *known* and you computed forward: given a fair coin, given an exponential lifetime with $\lambda = 0.02$, what will samples look like? That direction is probability. In this course the direction is reversed. The sample is what you know;

the population is what you want. Given 20 measured shafts, what can be said about μ for all 5,000—and with how much confidence? That reversed direction is *statistical inference*, and it is harder, because many different populations could have produced the sample you saw. Probability is still the engine underneath: it tells us how statistics behave, and that behavior is what lets us attach an honest uncertainty to every inference.

Random sample. Observations $X_1, X_2, \ldots, X_n$ that are independent and all drawn from the same distribution (i.i.d.).

For the machinery of inference to work, the sample must be a *random sample*: the observations $X_1, X_2, \ldots, X_n$ must be independent of one another and must all come from the same distribution. The abbreviation is i.i.d., independent and identically distributed. A convenience sample—the 20 shafts nearest the door, the containers from a single quiet week, the test scenarios your team happened to think of—breaks these conditions, and every formula built on them then reports uncertainty that is simply wrong. Before computing anything, always ask how the data were obtained; Section 1.3 gives that question its own vocabulary.

Safety lens. An autonomous vehicle's population is every driving situation it will ever face. That population can never be enumerated, so safety claims are always inferences from a finite test sample to an unbounded population. Whether the test scenarios are representative—an i.i.d.-like draw from real operating conditions—is the central question of AI verification and validation.

Example 1.2 — Testing a perception model

A safety team evaluates a pedestrian-detection model that will run on a delivery robot fleet. They replay 1,200 recorded street scenarios through the model and find that it fails to detect the pedestrian in 18 of them, a failure fraction of $18/1200 = 0.015$. Identify the population, the sample, the parameter of interest, and the statistic. Then state one question about this setup that belongs to probability (ISE 205) and one that belongs to statistical inference (this course).

Solution. The population is the collection of all street scenarios the fleet will encounter in operation—effectively infinite and impossible to enumerate. The sample is the 1,200 replayed scenarios. The parameter is p, the true failure probability of the model over the whole population of scenarios. The statistic is $\hat{p} = 18/1200 = 0.015$, the sample failure fraction.

A probability question assumes the population is known and computes forward: *if* the true failure rate is $p = 0.01$, what is the chance that a batch of 1,200 scenarios shows 18 or more failures? An inference question runs backward from the data: given that we observed $\hat{p} = 0.015$, what range of values of p is plausible, and can we be confident that p is below the safety requirement of 0.02? The second kind of question is what the chapters on

confidence intervals and hypothesis testing will answer.

Reference. Montgomery & Runger, 6th ed., Section 1-2.

1.3 *How Engineers Collect Data*

The value of a data set depends on how it was obtained, not only on how large it is. Montgomery and Runger distinguish three basic methods, and you should classify every data set you meet into one of them before analyzing it.

Retrospective study. A study that uses data already collected in the past for some other purpose, such as production logs or maintenance records.

A *retrospective study* uses historical data that already exist: production logs, maintenance records, last year's port records, the logged disengagements of a vehicle fleet. Retrospective data are cheap and plentiful, but they were recorded for other purposes. Key variables may be missing, recorded too coarsely, or logged only when something went wrong; transcription errors accumulate; and the conditions that produced the data may no longer hold.

Observational study. A study that observes and records a process as it runs, with as little interference as possible.

An *observational study* watches the process as it operates today, recording the variables of interest deliberately and carefully but without interfering. The engineer chooses what to measure and how precisely, which fixes most of the data-quality problems of retrospective data. What it does not fix is this: the inputs of the process change only as they happen to change on their own, so the interesting combinations may never occur, and variables that move together cannot be separated.

Designed experiment. A study in which the engineer makes deliberate, controlled changes to the inputs and observes the resulting outputs.

Confounding. Two variables changing together in the data, so their individual effects on the output cannot be separated.

In a *designed experiment* the engineer takes control: inputs are set deliberately to chosen levels, changes are made in a randomized order, and the outputs are observed. Deliberate, randomized change is what breaks *confounding*—the situation where two variables move together in the data so that their separate effects cannot be told apart. For this reason the designed experiment is the only one of the three methods that can establish cause and effect, and it is the gold standard whenever a causal claim must be defended.

Insight. Why only a designed experiment can establish causation. In retrospective and observational data, an input changed because something in the world changed it, and that same something may also have changed the output directly. The comparison is contaminated. In a designed experiment the input changes because the engineer's randomization said so—a cause disconnected from everything else in the system—so a systematic change in the output can only have come through the input.

Randomization is what converts ''these variables move together'' into ''this variable moves that one.''

Method	Who sets the inputs?	Cost	Main weakness
Retrospective	Nobody (data already exist)	Lowest	Missing or miscoded variables; conditions may have changed
Observational	The process itself	Moderate	Confounding; interesting input combinations may never occur
Designed experiment	The engineer, randomized	Highest	Cost and disruption of deliberate changes

Table 1.2. The three basic methods of collecting engineering data. Only the designed experiment supports cause-and-effect conclusions.

Startup lens. An A/B test is a designed experiment: visitors are randomly assigned to the old or new page, so a difference in conversion can be attributed to the page. Reading last quarter's churn out of the analytics database is retrospective—useful for spotting patterns, unreliable for proving what caused them.

Example 1.3 — Three plans for one startup question

A two-person startup wants to know whether a shorter signup form increases the fraction of visitors who create an account. They consider three plans. Classify each as retrospective, observational, or a designed experiment, and state which plans can support the causal claim ''the shorter form *increases* conversion.''

1. Pull last year's analytics: conversion was 3.1 % before a site redesign (long form) and 3.8 % after it (short form).
2. For the next month, record each visitor's device, arrival source, time on page, and whether they convert, changing nothing.
3. For the next month, randomly show each arriving visitor either the long form or the short form, and record conversions in each group.

Solution. Plan 1 is a retrospective study: the data already exist, collected for other purposes. It cannot support the causal claim, because the redesign changed many things at once—layout, colors, load time, even the season and the ad campaigns running at the time. The form length is confounded with all of them.

Plan 2 is an observational study: deliberate, careful measurement without interference. It will produce better-quality data than plan 1, but every visitor sees the same form, so it cannot compare the two forms at all.

Plan 3 is a designed experiment. Randomization ensures the two groups of visitors are alike in everything except the form they saw, so a systematic difference in conversion is attributable to the form. Only plan 3 supports the causal claim, and it is the plan the startup should run.

Engineering judgment. Designed experiments cost real production time: setting a furnace to a new temperature or re-routing a picking line disrupts operations. Engineering judgment decides which input ranges are safe to explore and how large a deliberate change the process can tolerate.

1.4 *Summarizing Data with Numbers*

Reference. Montgomery & Runger, 6th ed., Sections 6-1 and 6-2.

A raw list of observations $x_1, x_2, \ldots, x_n$ answers no engineering question by itself. The first act of statistics is to compress the list into a few numbers that describe its center and its spread.

Sample mean. The arithmetic average $\bar{x}$ of the observations; the standard measure of the center of a data set.

The *sample mean* is the arithmetic average of the observations:

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{1.1}$$

When we write the computed value from observed data we use the lowercase $\bar{x}$; the uppercase $\bar{X}$ denotes the same quantity viewed as a random variable, before the data are seen. The sample mean is the natural estimate of the population mean μ, and it has a physical interpretation: if each observation were a unit weight placed on a ruler at its value, $\bar{x}$ is the point where the ruler balances.

Sample variance. The average squared deviation from the sample mean, with divisor $n - 1$; the standard measure of spread.

The *sample variance* measures spread as the average squared deviation of the observations from their own mean:

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n - 1}. \tag{1.2}$$

Sample standard deviation. The positive square root of the sample variance; a spread measure in the same units as the data.

Squaring makes every deviation positive, so spread on both sides of the mean accumulates instead of canceling. The units of s^2 are the square of the data units (meters squared, hours squared), which is awkward to interpret, so in practice we report its square root.

The *sample standard deviation* is

$$S = \sqrt{S^2} = \sqrt{\frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n-1}}, \tag{1.3}$$

which is in the same units as the data and is the number to quote next to the mean: ''dwell time averaged 40.3 hours with a standard deviation of 14 hours'' is a complete first summary; the mean alone is not.

Insight. Why divide by $n-1$ and not n. The deviations $x_i - \bar{x}$ are measured from the sample's own mean, and they always sum exactly to zero. So once $n-1$ of them are known, the last one is forced: only $n-1$ of the deviations carry free information. Dividing by $n-1$ charges the spread to the number of independent deviations, and it makes S^2 an unbiased estimator of σ^2—on average, across many samples, S^2 neither systematically overstates nor understates the population variance. Dividing by n would systematically understate it, because the sample mean sits closer to its own data than μ does. The next chapter makes the word ''unbiased'' precise.

For hand computation, expanding the square in (1.2) gives an equivalent shortcut form that avoids computing each deviation separately:

$$S^2 = \frac{\sum_{i=1}^{n} X_i^2 - \dfrac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}}{n-1}. \tag{1.4}$$

Sample range. The difference between the largest and smallest observations, $r = \max x_i - \min x_i$.

Both forms give the same answer; use (1.2) when you want to see the deviations and (1.4) when you only have running totals of x_i and x_i^2, as a calculator does.

One more spread measure is worth a symbol. The *sample range* is

$$R = \max_i X_i - \min_i X_i, \tag{1.5}$$

the distance from the smallest observation to the largest. It is instant to compute and easy to explain, but it uses only two of the n observations and grows with sample size (more draws reach further into the tails), so it supplements s rather than replaces it.

Procedure 1.1 — Computing the summary statistics of a data set.

1. Record the sample size n and compute the total $\sum x_i$.
2. Compute the sample mean $\bar{x} = \sum x_i / n$ using (1.1).
3. Compute each deviation $x_i - \bar{x}$ and check that the deviations sum to zero (a built-in arithmetic check).
4. Square the deviations, sum them, and divide by $n - 1$ to get s^2 by (1.2); or apply the shortcut (1.4) to the totals $\sum x_i$ and $\sum x_i^2$.
5. Take the square root to get s by (1.3), and report $\bar{x}$ and s together, with units. Note the range (1.5) and the extremes if the application is safety-related.

The next example works Procedure 1.1 in full on a small data set that we will keep reusing in this chapter.

Example 1.4 — Miss distances of a perception model

A red team stress-tests the perception module of an autonomous shuttle. In eight staged crossing scenarios, the module detects the pedestrian late, and the recorded *miss distance*—how far from the intended stopping point the vehicle actually stopped—is measured. The data, in meters, are given in Table 1.3. Compute the sample mean, the sample variance, the sample standard deviation, and the sample range.

Scenario i	1	2	3	4	5	6	7	8
Miss distance x_i (m)	0.4	0.7	0.5	0.9	0.6	0.8	0.3	0.6

Table 1.3. Miss distances (meters) of a perception module in eight staged pedestrian-crossing scenarios.

Solution.

Step 1 — Size and total.
Here $n = 8$ and

$$\sum_{i=1}^{8} x_i = 0.4 + 0.7 + 0.5 + 0.9 + 0.6 + 0.8 + 0.3 + 0.6 = 4.8 \text{ m}.$$

Step 2 — Sample mean.
By (1.1),

$$\bar{x} = \frac{4.8}{8} = 0.6 \text{ m}.$$

Step 3 — Deviations, with the zero check.
Subtracting $\bar{x} = 0.6$ from each observation:

$$-0.2, \quad 0.1, \quad -0.1, \quad 0.3, \quad 0.0, \quad 0.2, \quad -0.3, \quad 0.0.$$

Their sum is 0, as it must be, so the arithmetic so far is consistent.

Step 4 — Sample variance.
Squaring and summing the deviations:

$$0.04 + 0.01 + 0.01 + 0.09 + 0.00 + 0.04 + 0.09 + 0.00 = 0.28 \text{ m}^2,$$

so by (1.2),

$$s^2 = \frac{0.28}{8 - 1} = \frac{0.28}{7} = 0.04 \text{ m}^2.$$

As a cross-check with the shortcut (1.4): $\sum x_i^2 = 0.16 + 0.49 + 0.25 + 0.81 + 0.36 + 0.64 + 0.09 + 0.36 = 3.16$, and $(\sum x_i)^2/n = 4.8^2/8 = 23.04/8 = 2.88$, so $s^2 = (3.16 - 2.88)/7 = 0.28/7 = 0.04 \text{ m}^2$. The two forms agree.

Step 5 — Standard deviation and range.
By (1.3), $s = \sqrt{0.04} = 0.2$ m. By (1.5), $r = 0.9 - 0.3 = 0.6$ m.

The summary is $\boxed{\bar{x} = 0.6 \text{ m}, \quad s = 0.2 \text{ m}}$, with $s^2 = 0.04 \text{ m}^2$ and range 0.6 m. In words: the module typically overshoots its intended stopping point by about 0.6 m, and scenario-to-scenario variation is about 0.2 m, with the worst observed case at 0.9 m.

Safety lens. For a safety engineer the mean is not the headline. The certification question is about the tail: how often does the miss distance exceed the clearance actually available? That is why Example 1.4 reports the extremes and the range alongside $\bar{x}$ and s, and why later chapters put confidence bounds on worst-case behavior rather than on averages alone.

Startup lens. Unit economics live and die on spread. Two products with the same average revenue per user are different businesses if one earns it steadily and the other from a handful of large accounts. Quote s next to $\bar{x}$ in your pitch numbers, and expect investors who know statistics to ask for it.

Reference. Montgomery & Runger, 6th ed., Sections 6-3, 6-4, and 6-5.

1.5 *Summarizing Data with Pictures*

Numerical summaries compress; pictures reveal. A good plot shows in one glance the center, the spread, the shape, and any observations that do not belong. This section presents the three plots used throughout the course.

Dot diagram. A plot of each observation as a dot along a measurement scale; the basic picture for small data sets.

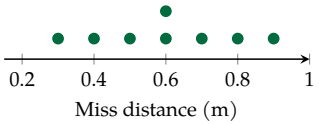

Figure 1.3. Dot diagram of the eight miss distances of Table 1.3. The stack at 0.6 marks the repeated value.

A *dot diagram* places one dot per observation along the measurement scale, stacking repeated values. Figure 1.3 shows the eight miss distances of Example 1.4: the eye immediately sees the center near 0.6, the spread from 0.3 to 0.9, and the roughly symmetric shape. For samples up to about 20 observations the dot diagram is the right first picture, and you should sketch one before computing anything—it is the fastest way to catch a typo like 9.0 entered for 0.9.

Histogram. A bar chart of frequencies over consecutive class intervals; the standard picture for the shape of larger data sets.

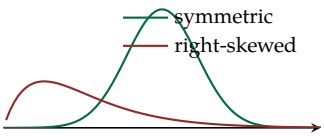

Figure 1.4. Two common histogram shapes. In the right-skewed shape the long tail pulls the mean above the median.

For larger samples, single dots blur together, and we group the data into consecutive class intervals of equal width, count the observations in each, and draw a bar per class whose height is the count. The result is a *histogram*. A workable rule of thumb is to use about $\sqrt{n}$ classes, adjusted to round class boundaries. The histogram trades detail for shape: individual values disappear, and center, spread, skewness, and multiple peaks appear.

Two shapes dominate engineering practice (Figure 1.4). Measurement-type data—dimensions, weights, scores—are often roughly symmetric and bell-shaped. Waiting-time-type data—dwell times, times to failure, response times—are often *right-skewed*: most values are moderate, a few are very large, and the long right tail pulls the mean above the typical value. Recognizing the shape matters, because several methods later in the course assume approximate symmetry.

Example 1.5 — Port container dwell times

A port operations team tracks how long import containers sit in the yard before leaving the terminal. For $n = 40$ containers, the dwell times were grouped into the frequency distribution of Table 1.4. Construct the histogram, describe the shape, and estimate the sample mean from the grouped data.

Dwell time class (h)	$[10,20)$	$[20,30)$	$[30,40)$	$[40,50)$	$[50,60)$	$[60,70)$	$[70,80)$
Class midpoint m_j (h)	15	25	35	45	55	65	75
Frequency f_j	3	7	11	9	6	3	1

Table 1.4. Frequency distribution of dwell times for 40 import containers.

Solution. The histogram is shown in Figure 1.5. The shape is single-peaked with its peak in the $[30, 40)$ class and a longer tail to the right: a mildly right-skewed shape, typical of waiting times. There are no gaps or isolated bars, so no observation demands investigation as an outlier.

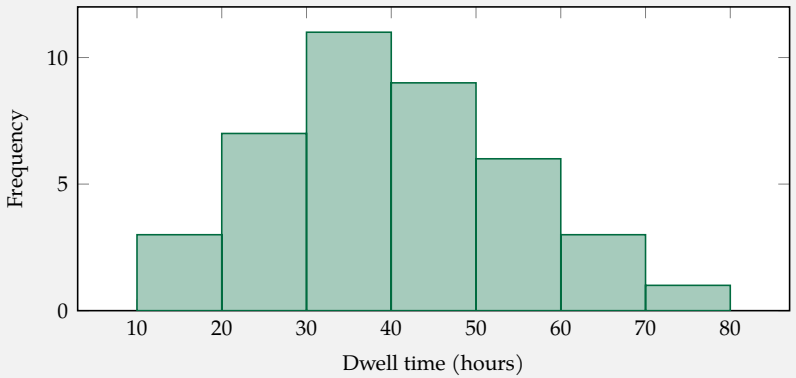

Figure 1.5. Histogram of the 40 container dwell times of Table 1.4. The longer right tail is typical of waiting-time data.

When only the grouped table is available, the sample mean is estimated by treating every observation in a class as if it sat at the class midpoint:

$$\bar{x} \;\approx\; \frac{\sum_j f_j\, m_j}{\sum_j f_j}, \tag{1.6}$$

where f_j is the frequency and m_j the midpoint of class j. Substituting the values of Table 1.4:

$$\sum_j f_j m_j = 3(15) + 7(25) + 11(35) + 9(45) + 6(55) + 3(65) + 1(75) = 1610,$$

so $\bar{x} \approx 1610/40 = 40.25$ h. The estimate is $\boxed{\bar{x} \approx 40.3 \text{ h}}$. Note that the peak of the histogram is in the thirties while the mean is above 40: the right tail has pulled the mean upward, exactly as Figure 1.4 predicts.

The third picture needs three more summaries, all based on sorting rather than averaging. The *sample median* is the middle value of the ordered data—half the observations below, half above. The first and third *quartiles* q_1 and q_3 do the same at the quarter marks, and the *interquartile range* $\mathrm{IQR} = q_3 - q_1$ measures the

Sample median. The middle value of the ordered data (or the average of the two middle values); half the observations lie on each side.

Quartiles. The values q_1 and q_3 that cut off the lowest and highest quarters of the ordered data; the median is q_2.

Interquartile range. The spread of the middle half of the data, $\mathrm{IQR} = q_3 - q_1$.

Box plot. A picture of the five-number summary: a box from q_1 to q_3 with the median inside, and whiskers to the extreme observations within 1.5 IQR of the box.

Outlier. An observation falling more than 1.5 IQR beyond the nearest quartile; plotted individually on a box plot.

spread of the middle half. Because these are positional, a single wild observation moves them barely at all, whereas it can drag $\bar{x}$ and inflate s badly. When a data set may contain outliers, median and IQR are the robust companions to mean and standard deviation.

A *box plot* draws these summaries: a box spanning q_1 to q_3 with a line at the median, and whiskers extending to the most extreme observations still within 1.5 IQR of the box. Any observation beyond the whiskers is plotted as an individual point and flagged as a potential *outlier*—not to be deleted, but to be investigated. Box plots earn their keep when several groups must be compared side by side, which is exactly the situation in the next example.

Example 1.6 — Comparing two onboarding flows

A startup pilots two onboarding flows for its scheduling product and records, for each pilot customer, the time from account creation to first successful booking. Flow A is the current design; flow B is a shorter design. The five-number summaries (in minutes) are given in Table 1.5, and the box plots in Figure 1.6. Compare the two flows: center, spread, and shape.

Flow	Min	q_1	Median	q_3	Max
A (current)	14	26	32	41	55
B (shorter)	11	20	24	30	42

Table 1.5. Five-number summaries of time to first booking (minutes) under two onboarding flows.

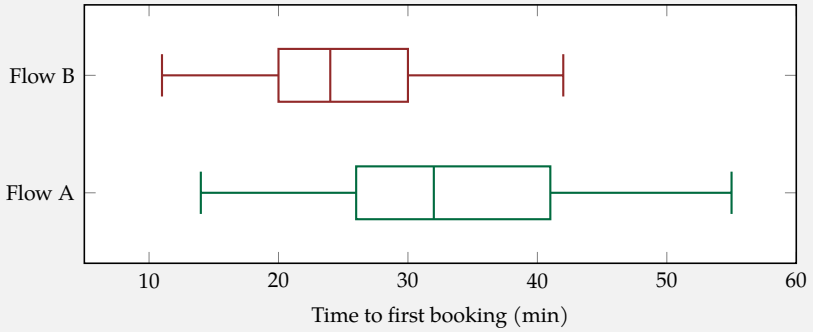

Figure 1.6. Comparative box plots of time to first booking under the two onboarding flows of Table 1.5.

Solution.

Center.

The median falls from 32 min under flow A to 24 min under flow B, so the typical pilot customer books 8 minutes sooner with the shorter flow.

Spread.

The interquartile ranges are $\text{IQR}_A = 41 - 26 = 15$ min and $\text{IQR}_B = 30 - 20 = 10$ min: flow B is not only faster in the middle but also more consistent.

Shape.

In both flows the upper whisker is longer than the lower one (A: $55 - 41 = 14$ above the box versus $26 - 14 = 12$ below; B: $42 - 30 = 12$ versus $20 - 11 = 9$), a mild right skew that is expected for time-to-complete data.

Every feature of the comparison favors flow B in this pilot. What the box plots cannot say is whether a difference of this size could have arisen by chance from the particular customers sampled—that question needs the two-sample methods of a later chapter, and this figure is the picture you should draw before running them.

Reference. Montgomery & Runger, 6th ed., Chapters 7–14.

1.6 *The Road Ahead: What This Course Covers*

Everything in this chapter has been description: honest summaries of the sample in hand. The rest of the course is inference: using those summaries to make defensible statements about the population, with the uncertainty quantified. Figure 1.7 shows the route.

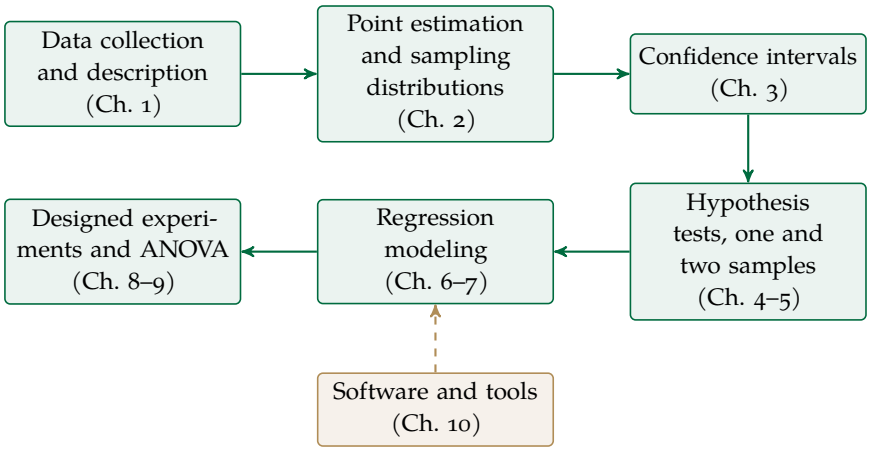

Figure 1.7. The route through this course: from describing a sample to estimating parameters, bounding them, testing claims about them, and modeling and designing with them. The software chapter supports every stage.

The route has five stages. First, *estimation* (Chapter 2): a statistic such as $\bar{X}$ is itself a random variable, and its sampling distribution—sharpened by the central limit theorem—tells us how far it typically lands from the parameter it estimates. Second, *confidence intervals* (Chapter 3): instead of a single guess for μ, σ^2, or p, we report an interval that captures the parameter with a stated confidence, so the uncertainty is part of the answer. Third, *hypothesis tests* (Chapters 4 and 5): a yes-or-no engineering claim—the failure rate is below 2 %, the new flow is faster than the old—is confronted with the data, under explicit control of the risk of a wrong answer. Fourth, *regression* (Chapters 6 and 7): relationships between variables are fitted, tested, and used for prediction. Fifth, *designed experiments* (Chapters 8 and 9): analysis of variance for comparing several treatments at once, then factorial designs that reveal interactions between factors. Chapter 10 maps every one of these methods to its spreadsheet and Python implementation, so nothing in the course stays hand-computation only.

Reference. Montgomery & Runger, 6th ed., Sections 1-1 and 1-2.

1.7 *Statistical Thinking, Step by Step*

Every problem in this course, and most statistical work you will do after it, follows the same skeleton: an engineering question, a population, a sample, a computation, an inference, and a decision. The skeleton is worth writing down as a numbered procedure, because the most common failure in practice is not bad

arithmetic—it is skipping straight to the computation with the population and the data-collection method never stated.

Procedure 1.2 — Reading a problem statistically (population → sample → inference).

1. **State the engineering question** in one sentence, including the decision that hangs on the answer.
2. **Identify the population** and the parameter (μ, σ^2, p, ...) that answers the question.
3. **Identify the sample**: what was observed, the sample size n, and how the data were collected (retrospective, observational, or designed experiment—and whether the observations are plausibly i.i.d.).
4. **Describe the sample**: compute the relevant statistics (Procedure 1.1) and draw the appropriate picture (dot diagram, histogram, or box plot).
5. **State the inference**: what the statistics say about the parameter, and with how much uncertainty. (Chapters 2–9 supply the tools for this step; until then, state it qualitatively.)
6. **Translate back to the decision**, checking that the data-collection method actually supports the kind of claim being made—in particular, that causal language is used only when a designed experiment stands behind it.

To see the procedure work, apply it to a guardrail problem. A team deploys a content guardrail in front of a customer-support chatbot, and the product owner asks: *is the guardrail blocking too many legitimate messages?* Step 1, the question: estimate the false-positive rate, because if it exceeds about 5 % the guardrail must be retuned before launch. Step 2, the population and parameter: all legitimate customer messages the deployed system will ever receive, with parameter p, the proportion wrongly blocked. Step 3, the sample: the team draws 400 messages at random from last month's human-verified legitimate traffic and replays them through the guardrail—a retrospective source, but randomly sampled from it,

and each message is scored independently; 14 are blocked. Step 4, description: $\hat{p} = 14/400 = 0.035$, and with a binary outcome the picture is just the count itself. Step 5, the inference, qualitatively for now: the sample estimate 3.5 % is below the 5 % threshold, but a different sample of 400 would have given a different $\hat{p}$, so we must ask whether values of p above 5 % are still plausible given these data—the confidence-interval methods of Chapter 3 answer exactly this. Step 6, the decision: no causal claim is involved, only an estimate, so the retrospective source is acceptable *if* last month's traffic resembles future traffic; if the product is changing quickly, that assumption itself should be flagged to the product owner.

Please work through this procedure explicitly, on paper, for the early homework problems. It feels slow for one week and then becomes the way you read every statistics problem for the rest of the course.

Chapter Summary

Engineering data vary, and statistics is the discipline of measuring that variability and reasoning under it. The population is the whole collection the question is about, described by fixed unknown parameters (μ, σ^2, p); the sample is what we observe, described by computable statistics ($\bar{x}$, s^2, $\hat{p}$). Probability (ISE 205) reasons from a known population to the behavior of samples; statistical inference (this course) reasons back from an observed sample to the unknown population. How data are collected bounds what they can prove: retrospective and observational data describe and suggest, but only a randomized designed experiment establishes cause and effect. A data set is described by its center ($\bar{x}$, median), its spread (s^2, s, range, IQR), and its picture (dot diagram for small samples, histogram for shape, box plot for comparisons). Procedure 1.1 computes the summaries; Procedure 1.2 organizes any statistical reading of an engineering problem, and both will be cited by number throughout the course.

Quantity	Formula	Equation
Sample mean	$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$	(1.1)
Sample variance	$S^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}$	(1.2)
Sample variance (shortcut)	$S^2 = \frac{\sum X_i^2-(\sum X_i)^2/n}{n-1}$	(1.4)
Sample standard deviation	$S = \sqrt{S^2}$	(1.3)
Sample range	$R = \max_i X_i - \min_i X_i$	(1.5)
Grouped-data mean	$\bar{x} \approx \sum_j f_j m_j / \sum_j f_j$	(1.6)

Table 1.6. Key formulas of Chapter 1 at a glance.

Exercises

Descriptive statistics

Exercise 1.1. A startup logs the number of new account signups on six consecutive weekdays: 12, 18, 15, 21, 14, 16. Compute the sample mean, the sample variance, and the sample standard deviation, following Procedure 1.1.

Exercise 1.2. A truck in a delivery fleet records the following fuel consumption on eight runs of the same route, in liters per 100 km: 32.1, 30.8, 33.4, 31.5, 32.9, 31.9, 32.6, 30.4. Compute $\bar{x}$, s^2, s, and the sample range r. Report each with its units.

Exercise 1.3. A calibration correction adds exactly 0.5 m to every one of the eight miss distances of Table 1.3. What happens to the sample mean and to the sample variance?

(A) Both increase by 0.5.

(B) The mean increases by 0.5; the variance is unchanged.

(C) The mean is unchanged; the variance increases by 0.25.

(D) Both are unchanged.

Types of studies

Exercise 1.4. Classify each data-collection plan as a retrospective study, an observational study, or a designed experiment, with one sentence of justification each.

1. An AV safety team pulls the last twelve months of logged disengagement reports from its fleet database to summarize disengagement frequency by city.
2. A warehouse engineer stands at the picking line for two weeks and records each picker's time per order, item weight, and shelf height, changing nothing about the operation.
3. A startup randomly assigns each arriving visitor to one of two prices for the same plan and records the purchase decisions.
4. A port analyst downloads last year's terminal records to compare dwell times before and after a gate-scheduling policy that was introduced in June.

Exercise 1.5. An observational study across 30 warehouses finds that warehouses using newer barcode scanners have clearly faster average picking times. A manager concludes that buying new scanners will speed up picking. Explain why the data do not establish this conclusion, name a plausible confounding variable, and describe a designed experiment that would settle the question.

Exam-style problems

Exercise 1.6. A guardrail service screens every prompt sent to a company chatbot. For eight production prompts sampled at random from one day's traffic, the added screening latency, in milliseconds, was:

$$42,\quad 38,\quad 51,\quad 45,\quad 39,\quad 48,\quad 44,\quad 41.$$

1. **[2 pt]** Identify the population and the sample, and state whether the value you will compute in part (b) is a parameter or a statistic.
2. **[3 pt]** Compute the sample mean latency.
3. **[4 pt]** Compute the sample variance and the sample standard deviation.

4. **[2 pt]** The eight prompts were sampled from existing logs. Which type of study is this, and does it support the claim ''the guardrail *causes* at most 50 ms of added latency''?

Exercise 1.7. A rail freight operator groups the loading times of $n = 40$ trains into five classes with midpoints $15, 25, 35, 45, 55$ minutes and frequencies $4, 10, 12, 8, 6$ respectively.

1. **[2 pt]** The data were extracted from the terminal's operations database. Classify the study type.
2. **[4 pt]** Estimate the sample mean loading time from the grouped data.
3. **[2 pt]** The frequencies rise to a single peak and fall away more slowly on the high side. Describe the shape of the histogram in one sentence, and state whether the median is above or below the mean.
4. **[2 pt]** Is the value computed in part (b) a parameter or a statistic? Justify in one sentence.

Assessing outside recommendations

Exercise 1.8. A founder asks an AI assistant to analyze five weekly signup conversion rates from a pricing pilot, in percent: 3.2, 4.0, 3.6, 4.4, 2.8. The assistant replies: ''Mean $= 3.6\,\%$. Variance $= \frac{(3.2-3.6)^2+(4.0-3.6)^2+(3.6-3.6)^2+(4.4-3.6)^2+(2.8-3.6)^2}{5} = 0.32$. Because the variance is small, the true conversion rate of the new pricing is $3.6\,\%$.'' Find every statistical error in this analysis, explain each, and give the corrected numbers.

Exercise 1.9. A vendor pitches a new perception stack to an AV operator with the following analysis: ''Your fleet logged 0.42 disengagements per 1,000 km last year. The three customers who installed our stack logged only 0.31 per 1,000 km over the same period. Installing our stack therefore reduces disengagements by 26 %.'' Using the vocabulary of Section 1.3 and Procedure 1.2, identify what kind of study this is, list at least two confounding variables that undermine the causal claim, and describe how a sound comparison would be designed.

In-Class Activity 1.1: First Data, First Decisions

Week 1 — Lecture 1

Work in teams of 3–4. Write your reasoning, not just the final number. One submission per team, all names on it.

Your team is the safety review board of a campus delivery-robot pilot. In its first week, the robot's emergency-stop system activated eight times. For six of the activations, the log records a usable measurement of the distance between the robot's final stopping point and the pedestrian, in meters:

$$0.5, \qquad 0.9, \qquad 0.4, \qquad 0.6, \qquad 0.8, \qquad 0.4.$$

Part 1 — What are we actually talking about? (5 min)

In one sentence each: what is the *population* here, what is the *sample*, and is the number your team will compute in Part 2 a parameter or a statistic? Be precise about what collection the campus administration actually cares about.

Part 2 — Summarize the evidence (5 min)

Compute the sample mean and the sample range of the six stopping distances, and sketch a quick dot diagram in the box. Then answer in one sentence: which single recorded value would you investigate first, and why?

t kind of study is this, and what can it prove? (*5 min*)

me from whatever activations happened to occur during normal operation. Classify this data-
od (retrospective, observational, or designed experiment). The vendor claims these numbers
always stops at least 0.4 m away.'' Write your team's one-sentence verdict on that claim, in the
/disagree because ...''

t populations, samples, or study types that is still unclear to your team (this seeds the next