

Foundation of Probability and Statistical Thinking for Health Data

3

Why Probability Comes Before Statistics

Every statistical statement about health data — a confidence interval on a vaccination coverage estimate, a p-value for a treatment comparison, a predicted probability of a positive malaria test — is a probability statement, and it can only be interpreted correctly by someone who understands the probability model that generated it. This chapter builds that foundation from first principles, using health examples throughout, before the book turns to inference (Chapter 7) and modelling (Chapters 8-15).

Random Variables and Distributions

A random variable maps the outcome of an uncertain process to a number. In health data, discrete random variables are pervasive: the number of confirmed malaria cases reported by a facility in a month, the number of children in a household under five, the result of a rapid diagnostic test (positive/negative). Continuous random variables include a patient's haemoglobin concentration, systolic blood pressure, or age at first antenatal visit.

Four distributions recur constantly in this book and deserve to be internalised early:

- The Bernoulli/Binomial distribution, for a single or repeated yes/no outcome such as a positive diagnostic test

result — the natural model behind logistic regression (Chapter 9) and test positivity rate calculations.

- The Poisson distribution, for counts of events occurring at some average rate over a fixed interval — the natural model for monthly disease case counts reported through HMIS, and the starting point for the Poisson generalised linear models used in Chapter 10's forecasting of NTD trends.
- The Normal (Gaussian) distribution, which, by the Central Limit Theorem, approximates the sampling distribution of many summary statistics (means, regression coefficients) even when the underlying data are not themselves normally distributed — this is why it underlies most classical confidence intervals and hypothesis tests.
- The Exponential and Weibull distributions, which model waiting times or time-to-event data such as time from HIV diagnosis to treatment initiation, or time to relapse — the foundation of the survival analysis methods in Chapter 10.

Expectation, Variance, and the Idea of a Statistical Model

The expectation (mean) of a random variable is its long-run average value; the variance measures the spread around that average. A statistical model formalises the claim that an observed variable is a realisation of a random variable with a particular distribution, whose parameters (for example, the Poisson rate of malaria cases per facility per month) we wish to estimate from data. Nearly every technique in this book — from a simple two-sample t-test to a deep neural network — can be understood as a way of estimating the parameters of an assumed (explicit or implicit) statistical model and then quantifying how confident we should be in that estimate.

Conditional Probability, Bayes' Theorem, and Diagnostic Testing

Conditional probability — the probability of an event given that another has occurred — is the single most important idea in this chapter for clinical and public health interpretation, because it underlies the difference between a diagnostic test's sensitivity and specificity on the one hand, and its positive predictive value (PPV) on the other. Bayes' theorem formalises the link:

$$P(\text{Disease} \mid \text{Positive test}) = [P(\text{Positive test} \mid \text{Disease}) \times P(\text{Disease})] / P(\text{Positive test})$$

Consider a rapid diagnostic test (RDT) for malaria with sensitivity 0.95 and specificity 0.90, deployed during the dry season in an area where true malaria prevalence among febrile presentations has fallen to 5%. Applying Bayes' theorem shows that the positive predictive value is only around 33%, meaning two out of every three positive RDT results in that setting are false positives. This is not a flaw in the test; it is a mathematical consequence of testing in a low-prevalence population, and it is precisely why national malaria case management guidelines emphasise pairing RDT results with clinical judgement and, where available, confirmatory microscopy or PCR — the three-way comparison that recurs as a running example throughout this book's diagnostics-related chapters.

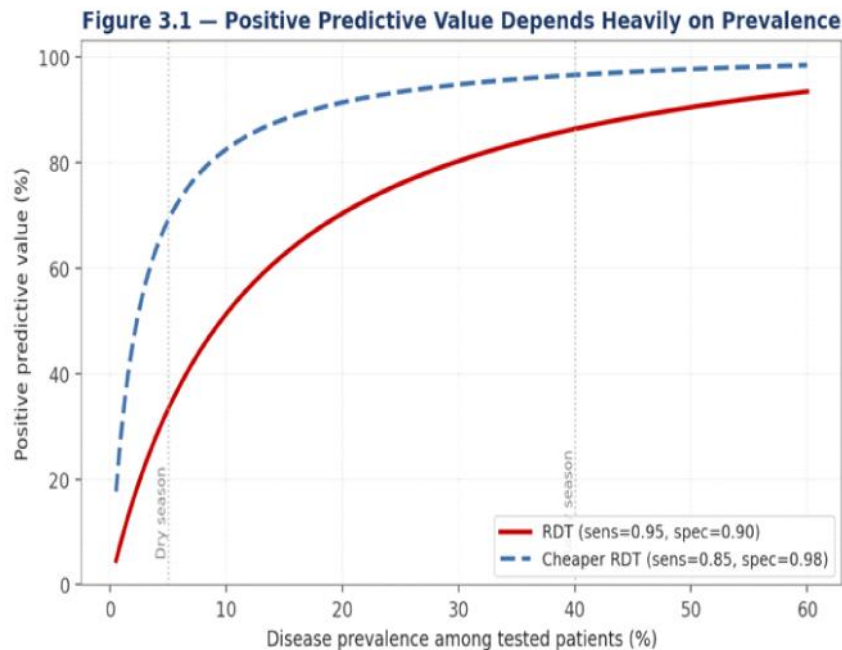


Figure 3.1 — Positive predictive value as a function of prevalence (Sections 3.4, 3.6, 3.9).

Sampling Distributions and the Logic of Uncertainty

A single facility's test positivity rate for a given month is a sample statistic — a snapshot from a random process — and it varies from what one would observe if every febrile patient nationwide could be tested. The sampling distribution describes how that statistic would vary if the process were repeated. Understanding this distinction is what separates a data scientist from someone who merely reads numbers off a dashboard: a facility reporting a TPR of 45% one month and 30% the next has not necessarily seen a real change in transmission; if patient volumes are small, that swing may be entirely consistent with sampling variability, and Chapter 7 develops the formal tools (standard errors, confidence intervals) to make this judgement rigorously rather than by eye.

Worked Example

Full Bayes' Theorem Calculation, Step by Step

It is worth working through the malaria RDT calculation from Section 3.4 in full arithmetic detail, because the habit of doing this calculation by hand at least once is what makes the result intuitive rather than merely memorised. Suppose 1,000 febrile patients are tested, and true prevalence among them is 5% (50 truly infected, 950 truly uninfected).

Group	True status	RDT result	Count
Sensitivity 0.95 applied to 50 infected	Infected	True positive	47.5 (≈ 48)
	Infected	False negative	2.5 (≈ 2)
Specificity 0.90 applied to 950 uninfected	Uninfected	True negative	855
	Uninfected	False positive	95

Total positive RDT results = 48 (true positive) + 95 (false positive) = 143. Positive predictive value = $48 / 143 \approx 33.6\%$, matching the figure quoted in Section 3.4. Repeating this calculation with a rainy-season prevalence of 40% instead of 5% (400 infected, 600 uninfected out of 1,000) gives true positives = 380, false positives = 60, PPV = $380 / 440 \approx 86.4\%$ — a dramatic illustration of how the same test, with the same sensitivity and specificity, delivers a very different practical reliability depending on the prevalence of disease in the population being tested.

PYTHON example

```
# Python: a small reusable function for Bayes' theorem applied to diagnostic testing
def positive_predictive_value(sensitivity, specificity, prevalence):
    true_pos = sensitivity * prevalence
    false_pos = (1 - specificity) * (1 - prevalence)
    return true_pos / (true_pos + false_pos)

for prevalence in [0.05, 0.20, 0.40]:
    ppv = positive_predictive_value(sensitivity=0.95, specificity=0.90, prevalence=prevalence)
    print(f"Prevalence {prevalence:.0%} -> PPV = {ppv:.1%}")
```

Simulating Distributions to Build Intuition

Simulation is one of the most effective ways to build genuine intuition for a probability distribution, and it will recur throughout this book (notably in Chapter 13's Monte Carlo methods). The code below simulates 10,000 draws from a Poisson distribution with a rate matching a moderate-transmission facility's average monthly case count, and compares the simulated distribution's mean and variance — which should be approximately equal, a defining property of the Poisson distribution exploited in Chapter 8's overdispersion diagnostics.

R example

```
# R: simulating a Poisson process and checking mean ≈ variance
set.seed(123)
simulated_cases <- rpois(n = 10000, lambda = 18) # average 18 cases/month

cat("Simulated mean:", mean(simulated_cases), "\n")
cat("Simulated variance:", var(simulated_cases), "\n")
hist(simulated_cases, breaks = 20, main = "Simulated Monthly Malaria Case Counts",
     xlab = "Cases per month", col = "steelblue")
```

Independence and the Danger of Assuming It Away

Many of the simplest formulas in introductory probability and statistics — for example, that the variance of a sum of random variables equals the sum of their variances — hold only under an assumption of independence, meaning knowing the value of one variable tells you nothing about another. This assumption is frequently violated in health data in ways that are easy to overlook: repeated measurements on the same patient are not independent of each other (Chapter 9's survival and Chapter 4's longitudinal data structures both hinge on this), and patients treated at the same facility are not independent of each other because they share facility-level influences such as staff training and equipment quality (the same clustering logic that motivates survey design effects in Chapter 7 and grouped cross-validation in Chapter 11). A recurring discipline this book asks of its readers is to explicitly

ask, for any new dataset, what the natural unit of independence actually is, rather than defaulting to treating every row as an independent observation simply because it occupies its own row in a spreadsheet.

Extended Case Study

The Base Rate Fallacy Beyond Diagnostic Testing

Section 3.4's diagnostic testing example is the most common illustration of the base rate fallacy in health data science, but the same underlying logical error recurs in a strikingly wide range of other applied settings, and it is worth working through a second, structurally identical example to consolidate the intuition. Suppose a machine-learning-based screening algorithm is developed to flag patient records at high risk of a rare adverse drug reaction, with a reported sensitivity of 90% and specificity of 95%, and suppose the true prevalence of this adverse reaction among all patients prescribed the drug is just 0.5% (a genuinely rare event, typical of many serious but uncommon adverse drug reactions). Applying exactly the same Bayes' theorem calculation as Section 3.6: out of 100,000 prescriptions, 500 will have a true adverse reaction (450 correctly flagged, 50 missed) and 99,500 will not (94,525 correctly cleared, 4,975 incorrectly flagged). The resulting positive predictive value is $450/(450+4,975)$ H" 8.3% — meaning that of every roughly 12 patient records the algorithm flags as high-risk, only about 1 will actually go on to experience the adverse reaction.

This is not a failure of the algorithm's underlying statistical performance (90% sensitivity and 95% specificity are both genuinely strong figures by ordinary standards) but an unavoidable mathematical consequence of screening for a rare event in a large population — precisely the same base-rate logic as the malaria RDT example, now applied to a pharmacovigilance context rather than a diagnostic test context. The practical implication for a

Nigerian pharmacovigilance or clinical decision-support system is that a low positive predictive value at this stage of algorithm deployment does not necessarily mean the tool should be discarded; it means the tool's output should be explicitly framed and used as a triage/prioritisation signal (worth a closer manual chart review) rather than as a near-certain diagnosis, and that any headline reporting of the algorithm's '90% sensitivity' figure to a non-technical stakeholder audience should always be accompanied by the corresponding positive predictive value at the relevant real-world prevalence, precisely to prevent the base rate fallacy from creating unrealistic expectations about how often a flagged case will turn out to be a true case.

PYTHON example

```
# Python: reusing the PPV function from Section 3.6 for a pharmacovigilance screening tool
def positive_predictive_value(sensitivity, specificity, prevalence):
    true_pos = sensitivity * prevalence
    false_pos = (1 - specificity) * (1 - prevalence)
    return true_pos / (true_pos + false_pos)

ppv = positive_predictive_value(sensitivity=0.90, specificity=0.95, prevalence=0.005)
print(f"PPV for rare adverse reaction screening: {ppv:.1%}") # approx. 8.3%
```

The Law of Large Numbers and Why it Reassures Us (Sometimes)

The law of large numbers states that as a sample grows, its average value converges toward the true population mean — the formal justification behind the everyday intuition that a larger sample gives a more reliable estimate. This law is frequently invoked, correctly, to argue that a large national survey's headline estimate is more trustworthy than a small facility's monthly figure. It is worth being precise, however, about exactly what the law does and does not guarantee: it guarantees convergence toward whatever the sample is actually drawing from, which is only the true population value if the sampling process itself is unbiased. A very large but systematically biased sample (for example, a very large

dataset drawn only from patients wealthy enough to afford private care) will converge, with great precision, toward the wrong number — precisely the point made about systematic bias in Section 1.5’s discussion of sample size versus bias, now given its formal statistical grounding.

Regression to the Mean

A Distinct Phenomenon from Bias

A statistical phenomenon easily confused with the systematic biases discussed elsewhere in this chapter, but conceptually distinct, is regression to the mean: an extreme observation (a facility with an unusually high TPR one month, a patient with an unusually high blood pressure reading at one visit) will, on average, be followed by a less extreme observation simply due to the role of chance in producing the original extreme value, even with no genuine underlying change and no systematic bias at all. This matters directly for programme evaluation: if a health programme deliberately targets the facilities or patients with the worst results in a baseline period for a special intervention, some of the improvement observed at follow-up would have occurred anyway, purely from regression to the mean, even if the intervention had no true effect whatsoever — a confound distinct from, and in addition to, the seasonal and reporting-completeness confounds discussed in Section 1.8’s IRS campaign case study, and one that specifically arises whenever a group is selected for study or intervention BECAUSE of an extreme baseline value.

Multiple Testing Revisited

Why This Isn’t Just a Chapter 7 Topic

Chapter 7 formally introduces multiple comparisons corrections, but the underlying probability logic belongs here in the foundations chapter: if a truly null effect has only a 5% chance of appearing ‘significant’ in any one test, the probability of at least one false

positive across 20 independent such tests is not 5% but 1 “ $(0.95)^{20}$ ” 64%, a dramatic escalation that follows directly from the basic probability rule for the union of independent events introduced in this chapter’s early sections. Internalising this calculation — rather than merely remembering the rule ‘correct for multiple comparisons’ as an isolated statistical convention — is what allows an analyst to recognise the same underlying risk in contexts far beyond formal hypothesis testing, including, for example, when scanning many exploratory plots (Chapter 5) or reviewing many candidate machine learning models (Chapter 11) for the one that looks best.

The Gambler’s Fallacy and Its Cousins in Health Decision-Making

The gambler’s fallacy — the mistaken belief that a run of one outcome makes the opposite outcome ‘due’ — has direct analogues in health decision-making that are worth naming explicitly. A clinician who has just seen three consecutive malaria-negative RDT results in busy fever season might, without realising it, become subtly more inclined to suspect the fourth febrile patient has malaria ‘because we’re due for a positive,’ when in fact (assuming a stable underlying prevalence) each test result remains statistically independent of the ones before it. This is a specific, avoidable instance of the independence violation discussed in Section 3.8, and recognising it by name — as a specific, well-documented cognitive bias with a name and a known correction — helps a practitioner catch the error in their own reasoning in real time, which is considerably harder to do when the bias remains an abstract, unnamed possibility.

A More Complete Simulation Function with Reproducibility Guarantees

The simulation in Section 3.7 relies on a global call to `set.seed()`, which works for a single script but becomes fragile once

simulation code is wrapped into a reusable function that might be called many times within a larger analysis, since each call would otherwise draw from wherever the global random state happened to be left by the previous call. A cleaner pattern accepts the seed as an explicit function argument, guaranteeing that any given call is exactly reproducible regardless of what other random operations preceded it elsewhere in the script — a small change that matters considerably once simulation is embedded inside a larger pipeline rather than run as a standalone script.

R example

```
# R: a self-contained, independently reproducible simulation function
simulate_case_counts <- function(n = 10000, lambda = 18, seed = 123) {
  old_seed <- if (exists(".Random.seed")) .Random.seed else NULL
  on.exit(if (!is.null(old_seed)) assign(".Random.seed", old_seed, envir = .GlobalEnv))
  set.seed(seed)
  rpois(n, lambda = lambda)
}

# Two independent calls now give identical, reproducible results
identical(simulate_case_counts(seed = 1), simulate_case_counts(seed = 1)) # TRUE
```

PYTHON example

```
# Python: the equivalent pattern using a local Generator instance
import numpy as np

def simulate_case_counts(n=10000, lam=18, seed=123):
    rng = np.random.default_rng(seed) # local, not global – safe to call repeatedly
    return rng.poisson(lam=lam, size=n)

assert (simulate_case_counts(seed=1) == simulate_case_counts(seed=1)).all()
```

Chapter Summary

- Probability distributions (Bernoulli/Binomial, Poisson, Normal, Exponential/Weibull) provide the generative models underlying nearly every statistical method in this book.
- A statistical model is a claim about which distribution, and which parameters, generated the observed data.

- Bayes' theorem explains why diagnostic test positive predictive value depends heavily on underlying prevalence — a critical concept for malaria and NTD case management.
- Sampling variability means small-facility indicators can swing substantially by chance alone; this must be distinguished from a real change in disease transmission.
- Working through a Bayes' theorem calculation by hand, and simulating a distribution in code, both build the kind of intuition that formula memorisation alone does not provide.

Mini-Project

Mini-Project 3: A Prevalence-Sensitive Diagnostic Calculator

Build a small, reusable tool (in R or Python) that reports the positive and negative predictive value of any diagnostic test across a range of prevalence values, and use it to write a one-paragraph clinical recommendation.

- Step 1.** Extend the Python function in Section 3.6 (or write an equivalent R function) to also compute negative predictive value.
- Step 2.** Run the function across at least five prevalence values spanning a realistic range for a disease of your choice (e.g., 1% to 50%).
- Step 3.** Present the results as a table or simple line plot of PPV and NPV against prevalence.
- Step 4.** Identify the approximate prevalence at which PPV drops below 50% for your chosen test's sensitivity and specificity.
- Step 5.** Write a short paragraph advising a clinician on how to interpret a positive test result differently in a high- versus low-prevalence season.

Deliverable: *Working code, one plot or table, and a short written clinical recommendation.*

End-of-Chapter Exercises

1. A malaria RDT has sensitivity 0.95 and specificity 0.90. If prevalence among tested patients is 40% (peak transmission season), compute the positive predictive value and compare it to the 5%-prevalence example in Section 3.4.

2. Explain why monthly case counts from a single small health facility are better modelled as Poisson than as Normal.
3. A facility's TPR rises from 28% to 41% between two consecutive months with only 20 tests performed each month. Discuss, without doing a formal test, why this alone should not be read as evidence of a transmission increase.
4. Using the step-by-step method in Section 3.6, calculate the false negative rate (number of infected people who test negative) when prevalence is 40%, sensitivity is 0.95, and 1,000 patients are tested.
5. A new, cheaper RDT has sensitivity 0.85 and specificity 0.98. Using the Python function in Section 3.6, compute its PPV at 5% prevalence and compare it to the original test's PPV. Which property (sensitivity or specificity) matters more at low prevalence, and why?
6. Modify the R simulation in Section 3.7 to use a Poisson rate of $\lambda = 3$ (a low-transmission facility). Run it and comment on whether the mean and variance are still approximately equal.
7. In your own words, explain the difference between $P(\text{Disease} \mid \text{Positive test})$ and $P(\text{Positive test} \mid \text{Disease})$, and why confusing the two is a common and consequential error in interpreting diagnostic tests.
8. Give one example, other than those in Section 3.8, of health data where treating each row as independent would be incorrect, and explain the true unit of independence.
9. Explain why the variance of a sum of two correlated random variables is not simply the sum of their individual variances.
10. A colleague computes a standard error assuming independence for data that is actually clustered by facility. Will this standard error typically be too large or too small? Explain why.

11. Using the pharmacovigilance example in Section 3.9, recompute the positive predictive value if the true prevalence of the adverse reaction were 2% rather than 0.5%, keeping sensitivity and specificity unchanged.
12. Explain why a 90% sensitivity, 95% specificity screening tool can still have a very low positive predictive value, without this being a sign the tool is poorly built.
13. Propose one way a pharmacovigilance team could communicate the tool's output to clinical staff so that a 'flagged' record is not mistaken for a confirmed adverse reaction.
14. In Section 3.9, explain why increasing specificity would generally help the tool's positive predictive value more than a similar increase in sensitivity, given the very low base rate.
15. Give one other Nigerian health screening context (besides malaria RDTs and pharmacovigilance) where the base rate fallacy could mislead a non-technical audience, and briefly explain why.
16. Explain, using Section 3.10, why a very large sample drawn only from urban, private-hospital patients would not necessarily give a trustworthy estimate of national child mortality.
17. What does the law of large numbers guarantee, precisely, and what does it NOT guarantee, according to Section 3.10?
18. Give an example of a Nigerian health dataset that could be very large in sample size but still systematically biased in the way Section 3.10 describes.
19. Why is 'we have a huge sample size' sometimes used inappropriately to defend a conclusion, in light of Section 3.10?

20. Explain briefly how you would investigate whether a large dataset's size is actually protecting it from bias, rather than simply assuming it does.
21. Using Section 3.11, explain why some improvement would be expected at follow-up even if an intervention targeting the worst-performing facilities had no true effect at all.
22. How is regression to the mean different from the seasonal and reporting-completeness confounds discussed in Section 1.8?
23. Propose a study design feature that could help distinguish a genuine intervention effect from regression to the mean.
24. Why does selecting facilities specifically BECAUSE of an extreme baseline value make regression to the mean especially likely to occur?
25. Explain, in your own words, why regression to the mean is a genuine statistical phenomenon and not evidence of dishonest reporting.
26. Explain, in a few sentences, why this chapter argues probability must be understood before statistical inference can be interpreted correctly.
27. Summarise the common thread linking the diagnostic testing example (Section 3.4), the pharmacovigilance example (Section 3.9), and the law of large numbers discussion (Section 3.10).
28. Why does this chapter emphasise working through calculations by hand at least once, rather than relying solely on software output?
29. Propose a scenario from your own field where the base rate fallacy could mislead a well-intentioned analyst.
30. What single probability concept from this chapter do you find most useful for your own work, and why?

31. Describe, using Section 3.13, a scenario in clinical decision-making where the gambler's fallacy could subtly bias a diagnosis.
32. Why is naming a cognitive bias explicitly (rather than leaving it as an abstract possibility) helpful for catching it in real time?
33. Explain why three consecutive negative test results do not make a positive result 'due' on the fourth test, assuming a stable underlying prevalence.
34. Propose one other context (outside diagnostic testing) where a health professional might unconsciously apply the gambler's fallacy.
35. How does the gambler's fallacy relate to, but differ from, the base rate fallacy discussed earlier in this chapter (Section 3.9)?

Part 2

Data Wrangling and Exploratory Analysis

