DATA SCIENCE & ML

Statistical Foundations for Data Scientists

Inference, experimentation and A/B testing done right

Houssam Kodad

PDF · DATAFORGE BOOKS

Questions about this sample or the full edition: **support@dataforgebooks.com**

CONTENTS

Table of Contents

CHAPTER 01

Thinking in Distributions

Many data scientists can train a model but hesitate when asked the more basic question: is this result real, or could it be noise? That question is the heart of statistics, and answering it well is what lets you make confident decisions instead of confident-sounding guesses. This book rebuilds the statistical foundations that matter day to day, with intuition and correct practice rather than proofs.

This chapter starts where statistics starts: with the idea that we observe samples but care about populations, that every estimate carries uncertainty, and that the size of that uncertainty is something we can actually quantify. We meet variance and the standard error, and the central limit theorem — the quiet engine that makes most of the methods in the book work at all.

The aim is a way of thinking, not a formula sheet. By the end of the book you will design an experiment that survives scrutiny, read a p-value without the common misinterpretations, and communicate uncertainty to people who do not want a statistics lecture. It begins with learning to see numbers as estimates with error bars rather than facts.

Populations and Samples

Almost everything in statistics flows from a single distinction: the population is what you care about, and the sample is what you can actually see. You rarely measure an entire population, so you estimate its properties from a sample — and because a different sample would have given a slightly different answer, every estimate is uncertain. Internalising this turns each number you report from a fact into an estimate with a margin around it.

This shift has immediate practical force. A conversion rate of 4.2% measured from a sample is not the truth; it is an estimate of the truth, and how much you should trust it depends on how much data stands behind it. Statistics is, at bottom, the disciplined practice of quantifying that margin honestly — and refusing to let a precise-looking decimal masquerade as certainty it has not earned.

Variance and the Standard Error

Two quantities are easy to confuse and important to separate. Variance measures how spread out individual values are. The standard error measures something more useful: how much your estimate of a quantity would wobble if you repeated the whole sampling process. It is the standard error, not the variance, that tells you how precise your headline number is.

The standard error has a property with hard practical consequences: it shrinks as the sample grows, but only with the square root of the sample size. That means halving your uncertainty requires four times the data, not twice. This single relationship explains why early results are so noisy, why doubling a small experiment barely helps, and why so much of experimental design is a negotiation with a stubborn square-root curve.

```
import numpy as np

sample = np.random.binomial(1, 0.042, size=2000)    # observed conversions
p_hat  = sample.mean()
se     = np.sqrt(p_hat * (1 - p_hat) / len(sample)) # standard error
ci     = (p_hat - 1.96 * se, p_hat + 1.96 * se)      # ~95% interval
```

The Central Limit Theorem

The central limit theorem is the reason the normal distribution appears everywhere, and it is more surprising than it first sounds. It says that the average of many independent observations is approximately normally distributed, almost regardless of the shape of the underlying data. Your data can be wildly skewed; the averages you compute from it will still be well behaved.

This is the free lunch that licenses much of what follows. Because sample means tend toward normality, you can build confidence intervals and run tests using the normal distribution even when the raw data looks nothing like a bell curve, provided your sample is reasonably large and your observations roughly independent. Understanding why this works keeps you from misapplying it when those conditions fail.

Confidence Intervals

A confidence interval expresses an estimate together with its uncertainty, and it is almost always more useful than a point estimate alone. Rather than claiming the conversion rate is 4.2%, you report that it plausibly lies between 3.4% and 5.0% — a statement that immediately tells the reader how much to trust the headline and whether two numbers are meaningfully different.

The interpretation deserves care, because it is widely mangled. A 95% confidence interval does not mean there is a 95% probability the true value lies inside this particular interval; it means the procedure produces intervals that capture the truth 95% of the time. The distinction sounds pedantic but matters when communicating results, and we get it right so that you can too.

Hypothesis Testing Without Myths

Hypothesis testing is the formal machinery for asking whether an observed effect is more than noise, and it is surrounded by more misconceptions than any other topic in applied statistics. The core idea is modest: assume there is no effect, compute how surprising your data would be under that assumption, and if it is surprising enough, conclude something is going on.

The p-value is where the myths cluster. It is the probability of data at least this extreme if the null hypothesis were true — and it is emphatically not the probability that the null is true, nor the probability your result is a fluke, nor a measure of effect size. We spend real effort dismantling these misreadings, because they lead competent people to confident wrong conclusions every day.

Designing an A/B Test

The most common application of statistics in industry is the A/B test, and most of its value or failure is decided before any data arrives, at the design stage. You must state the hypothesis and the metric in advance, and you must calculate the sample size needed to detect an effect worth caring about — because a test too small to see a real effect wastes everyone's time and produces a misleading null result.

This forward planning is what gives a test its credibility. Deciding the metric, the minimum effect of interest, and the required sample size up front protects you from the temptation to find a winner after the fact in whichever slice happens to look good. We treat power and sample-size calculation as the non-negotiable first step of every experiment, not an afterthought.

The Ways Experiments Lie

Even a well-designed experiment can mislead you, and the failure modes are systematic enough to name. Peeking at results and stopping the moment they look significant inflates false positives dramatically. Testing many metrics or segments and celebrating whichever crosses the line is the multiple-comparisons trap. Novelty effects make a change look better than it is simply because it is new.

Knowing these traps is half of avoiding them, and the other half is procedural discipline — fixing the analysis plan in advance, correcting for multiple comparisons, running tests long enough to outlast novelty. The book devotes real attention to them because the most dangerous statistical errors are not arithmetic mistakes but valid calculations applied to a quietly biased process. Guarding against that is the mark of a trustworthy analyst.

This is a free sample

You've reached the end of the sample chapter.

Get the complete book — every chapter, fully worked — at **dataforgebooks.com**.

FULL EDITION · 196 PAGES · PDF

Read the full title at **dataforgebooks.com**

Questions? support@dataforgebooks.com