We begin by framing our question as two opposing statements: the null hypothesis, which assumes no real effect or difference, and the alternative hypothesis, which captures the pattern we suspect. Next, we collect our data—ideally from a randomized experiment or a well-designed observational study—and choose an appropriate statistical test. We then compute a test statistic that summarizes how far our observed result deviates from what the null hypothesis predicts, and convert that into a p-value, which tells us the probability of seeing data at least this extreme if the null were true. By comparing the p-value to a predetermined threshold (commonly 0.05), we decide whether to reject the null hypothesis: if the p-value is smaller, we call the finding “statistically significant,” meaning it is unlikely to be due to random chance alone. Finally, we report both the p-value and an effect size along with confidence intervals, so that we convey not just whether the effect exists but also how large and precise it is.

The Central Limit Theorem tells us that when we draw sufficiently large random samples from any population the distribution of the sample means will be approximately normal, centering on the true population mean and having a spread equal to the population’s standard deviation divided by the square root of the sample size. This theorem is crucial because it underpins most of our familiar inferential tools—confidence intervals, hypothesis tests, control charts—and lets us use the Normal distribution as a universal approximation.

We define statistical power as the probability that we correctly detect a real effect—in other words, the chance we reject the null hypothesis when the alternative hypothesis is true. Mathematically, power = 1 – β, where β is the risk of a Type II error (failing to detect an effect that actually exists). Power depends on our chosen significance level (α), the true effect size, the variability in the data, and the sample size: larger samples, bigger effects, lower noise, or a higher α all boost power.

We begin by explicitly identifying the kinds of bias that can creep into our work—selection bias, measurement bias, confirmation bias, confounding, and so on—and then build our study design and analysis plan to counteract each one. For selection bias, we use random sampling (or, in experiments, random assignment) and, when necessary, apply weighting or matching techniques so our sample truly represents the population we care about. To guard against measurement bias, we standardize protocols, calibrate instruments, and, where feasible, blind both participants and analysts to treatment assignments so that expectations can’t influence observations. We combat confirmation and analytic bias by pre-registering our hypotheses and analysis scripts, adopting clear decision rules for data cleaning and outlier handling, and by splitting data into training and hold-out sets (or using cross-validation) to ensure our findings generalize. Finally, we report our methods, code, and sensitivity analyses transparently—so that others can reproduce our results and verify that our conclusions aren’t artifacts of hidden biases.

We call a variable a confounder when it is related both to the treatment (or exposure) we’re studying and to the outcome we care about, such that its influence can muddy the true relationship between them. In other words, a confounding variable creates a spurious association by acting as a hidden “third wheel” that drives both the independent and dependent variables.

We conduct A/B testing as a randomized experiment in which we split our users (or units) into at least two groups—A (the control) and B (the variant)—and expose each group to a different version of a webpage, feature, or treatment. By measuring the same key metric (click-through rate, conversion, time on page, etc.) in both groups and then applying a statistical test (for example, a t-test or chi-square test), we can compute a p-value to assess whether any observed difference is unlikely to have arisen by chance. If the result crosses our predefined significance threshold (often 0.05), we conclude that one version truly outperforms the other.

We compute a confidence interval around an estimate (like a sample mean or proportion) to capture a range of values that, under repeated sampling, would contain the true population parameter a specified proportion of the time. Concretely, we take our point estimate, add and subtract a “margin of error”. Saying “95% confidence” means that if we drew many independent samples and built a 95% interval each time, about 95% of those intervals would cover the true mean—it does not mean there’s a 95% chance the particular interval from our data contains the truth.