

Hypothesis testing revisited

Hicham Janati

Fall 2026

Every chapter so far has ended the same way: a model was written down and its parameters were estimated, and in chapter 1 the estimates came with a measure of their own uncertainty. This chapter turns that machinery around. Instead of asking *what value the data suggest for a parameter*, it asks whether the data support or undermine a *claim*: that a player is better than another, that a coefficient is zero, that a die is fair. You have met the answer’s vocabulary in first year: null hypothesis, p-value, level, rejection. The purpose here is to rebuild that vocabulary from its starting point, a single probability computation performed three centuries ago, and to assemble, step by step, the reasoning ladder that every statistical test follows, including the ones in the chapters ahead. The running example is two friends and a football video game.

1 Eighty-two years of boys

The London Bills of Mortality recorded yearly christenings, boys and girls separately, through the seventeenth and early eighteenth centuries. Over the 82 consecutive years from 1629 to 1710, every single year recorded more boys christened than girls [1]. What is the probability of that?

The reflex is to compute. Treat each year as a coin flip (boys or girls in the majority, each with probability $1/2$) and treat the years as independent. Then

$$\mathbb{P}(82 \text{ boy years out of } 82) = \left(\frac{1}{2}\right)^{82} \approx 2.1 \times 10^{-25}. \quad (1)$$

The computation is correct, and it does not answer the question that was asked. It answers a different one: *if* each year were a fair, independent coin flip, what would the chance of the record be? The two assumptions were loaded silently. Where did the $1/2$ come from? Nothing in the records says that boy-majority and girl-majority years are equally likely; we supplied that. Who said the years are independent? We did, without noticing. The number 2.1×10^{-25} is not a fact about London; it is a fact about a *model* of London that we chose in the act of computing.

That observation leaves exactly two readings. Either an event of probability 10^{-25} happened (an event so rare that no human institution has ever recorded one), or at least one of the loaded assumptions is false. John Arbuthnot, the physician who published this computation in 1710, took the second reading: the record is overwhelming evidence *against* the fair-coin model of births. This inversion is the founding move of statistical testing. A probability computed *under* an assumption is turned into evidence *about* the assumption: assume, compute the probability of what was observed, and let a tiny result testify against the assumption package.

The computation has a limit that is worth stating before the machinery makes it precise. It cannot say *which* part of the package fails. Perhaps boys are genuinely more probable at birth (they are, slightly); perhaps the years are not independent; the number 2.1×10^{-25} indicts the package as a whole and is silent about the parts. Arbuthnot himself overreached on this point: he read the rejection as proof of a specific cause (in his case, divine design), when the calculation licenses only “not chance, as chance was modeled here.” Keeping honest books (what is assumed, what is claimed, and what the computed number does and does not say) is exactly what the formal machinery of this chapter is for.

2 Who’s the better gamer ?

This section turns Arbuthnot’s inversion into a method, on a dispute closer to home, in five steps that the rest of the chapter extends: pose the question, translate it into hypotheses, compress the evidence into a statistic, derive that statistic’s distribution under the null, and turn a tail probability into a decision with known error rates.

2.1 A question about a game

Karim and Yassine, two students of this course, have played FIFA against each other all semester. Karim has won 14 of the last 20 games and claims to be, simply, the better player. Yassine’s answer is one word: luck. The dispute is real, and it is decidable, but not by staring at the score. Fourteen wins out of twenty is perfectly possible between equal players, and it is also what a genuinely better player might produce. The question that data can answer is Arbuthnot’s question, asked deliberately this time: *if the two were equally good, how surprising would 14 out of 20 be?*

To ask it cleanly, give the assumption a name instead of smuggling it in. Let θ denote the probability that Karim wins a single game, and model the 20 games as independent, each won by Karim with probability θ .¹ The dispute becomes a statement about θ :

$$H_0 : \theta = \frac{1}{2} \quad \text{against} \quad H_1 : \theta > \frac{1}{2}. \quad (2)$$

Definition 2.1 (Null and alternative hypotheses)

A statistical hypothesis is a statement about the parameter of a model. A test opposes a *null hypothesis* H_0 , the statement to be challenged, to an *alternative hypothesis* H_1 , the claim that would replace it. In (2) the null is *simple*: it specifies the parameter completely. A null covering a whole set of values is *composite* (section 2.5). The alternative here is *one-sided*: it points in one direction.

Two remarks on the casting of roles. First, why exact equality, when Yassine’s real claim is broader: Karim is *not better*, that is $\theta \leq 1/2$? Because computation needs a pinned-down model, and neither the full claim nor the alternative provides one: “not better” allows every θ up to $1/2$, “better” every θ above it. The boundary value $\theta = 1/2$ is the one point that does pin the model down (as the implicit $1/2$ per year did in section 1), and it is the right point to single out: of all the values “not better” allows, $1/2$ produces lopsided scores most easily, so a rejection calibrated there rejects the broader claim as a whole (sections 2.5 and 3.3 make this precise). Second, the two hypotheses will not be treated symmetrically. Everything that follows measures evidence *against* H_0 ; the precise sense in which the null can be rejected but never accepted is the destination of section 2.5.

2.2 A statistic and its null distribution

The data are twenty game results. To weigh them, they must first be compressed into a single number that measures the discrepancy from H_0 in the direction of the claim. Such a number is called a *test statistic*. Here the choice is obvious: the number of wins. (Arbuthnot’s statistic was the number of boy-majority years, 82 out of 82, chosen without ever being named.) A good statistic has two properties: it is computable from the data alone (no unknown parameter in its formula), and its distribution under H_0 is known, or at least obtainable.

For the number of wins the null distribution comes from one line. Write $S = \sum_{i=1}^{20} W_i$, where W_i indicates a Karim win in game i . Under H_0 the W_i are independent Bernoulli variables with

¹The letter θ , not p , on purpose: throughout this chapter p is reserved for the p-value, defined in section 2.3.

parameter $1/2$, so

$$\text{under } H_0 : S \sim \text{Bin}\left(20, \frac{1}{2}\right), \quad \mathbb{P}(S = k) = \binom{20}{k} \left(\frac{1}{2}\right)^{20}. \quad (3)$$

This distribution, the law of the statistic in the world where the null is true, is called the *null distribution*, and it is the measuring device of the whole chapter: it says which values of S are ordinary under H_0 and which are remarkable.

There is a second route to the same object. Instead of the formula, *simulate* the world of H_0 : make a computer play 100 000 fake seasons of 20 fair coin flips each and histogram the number of wins. Figure 1 shows the result on top of the exact distribution (3): the two agree, as they must. The formula is faster and exact; the simulation needs nothing but the ability to generate data under the null, a route that stays open even for statistics whose null distribution admits no usable formula.

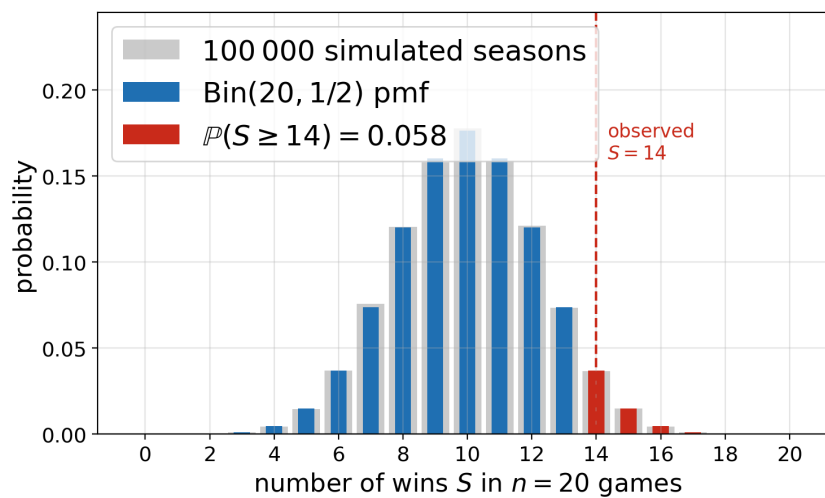


Figure 1: The null distribution of the number of wins S in 20 games: the exact $\text{Bin}(20, 1/2)$ probabilities (blue) and the histogram of 100 000 simulated fair seasons (gray) agree. The observed $S = 14$ sits in the right tail; the red bars are the outcomes at least as favorable to Karim, of total probability 0.058.

Nothing in this section is new probability. A model, a statistic, a sampling distribution under the model: this is chapter 1’s machinery in miniature, with one Bernoulli parameter in place of a mean vector and a covariance. The new idea is the direction of use. Estimation read the sampling distribution forward, to attach uncertainty to $\hat{\theta}$; testing reads it backward, as a scale on which the observed value of the statistic is judged for surprise.

2.3 The p-value

The observed statistic is $S = 14$. On the null distribution (3), how remarkable is it? The relevant region is not the single value 14 but everything at least as favorable to the claim, because the question is directional: the alternative in (2) points toward large S . Summing the tail of (3),

$$\mathbb{P}(S \geq 14) = 0.058 :$$

about one twenty-game season in seventeen between *equal* players produces a score at least this lopsided. This tail probability is the test’s summary of the evidence and has a name.

Definition 2.2 (p-value)

The *p-value* of an observed statistic s_{obs} is the probability, computed under the null distribution, of a statistic at least as extreme as the one observed, where “extreme” means *in the direction of the alternative*:

$$p = \mathbb{P}_{H_0}(S \geq s_{\text{obs}}) \quad (\text{here, for } H_1 : \theta > \tfrac{1}{2}). \quad (4)$$

A small p-value means: a world where H_0 holds rarely produces data this favorable to H_1 .

Set the two computations of this chapter side by side. Arbuthnot’s p-value, 2.1×10^{-25} , leaves no doubt. Karim’s, 0.058, is not small enough to rule out luck and not large enough to dismiss the claim. Deciding in such cases requires a rule fixed in advance.

2.4 The level and the threshold

The rule is fixed *before* looking at the data. Choose a *level* α (the convention in most fields is 5%) and decide as follows: reject H_0 when the p-value is at most α . Equivalently, since $p \leq \alpha$ exactly when the statistic falls beyond the corresponding quantile of the null distribution, the rule can be written on the scale of S itself:

$$\text{reject } H_0 \iff p \leq \alpha \iff S > q_{1-\alpha}, \quad (5)$$

where $q_{1-\alpha}$ is the $(1-\alpha)$ -quantile of the null distribution (the smallest q with $\mathbb{P}_{H_0}(S \leq q) \geq 1-\alpha$), read off in the direction the alternative points to; the statistic must strictly exceed it. The threshold is not a new object: it is a quantile of the null law, and the alternative decides *which* quantile (upper here, lower or two-sided for other alternatives; see section 3). For $\text{Bin}(20, 1/2)$ at $\alpha = 5\%$, the quantile is $q_{0.95} = 14$, and the tail probabilities are $\mathbb{P}(S \geq 14) = 0.058$ and $\mathbb{P}(S \geq 15) = 0.021$: the rejection region is $S > 14$, that is $S \geq 15$. Karim’s 14 falls one game short, and the test does not reject.

One feature of this example repays attention. No rejection region of the form $S \geq k$ has probability exactly 5% under $\text{Bin}(20, 1/2)$: the choice is between 5.8% (too lax) and 2.1% (the real size of the test actually used). A discrete statistic offers only a staircase of achievable levels, and the convention is to step down, so the guaranteed level is conservative. Exact-5% rejection regions are a luxury of continuous statistics.

2.5 Type I and Type II Errors and the Power of a Test

Return to the rule of section 2.4: reject when $S \geq 15$. Its properties can be computed before they are named. If the players are equal, $\theta = 0.5$, the rule rejects with probability $\mathbb{P}_{0.5}(S \geq 15) = 0.021$. If Karim actually wins 60% of his games, it rejects with probability $\mathbb{P}_{0.6}(S \geq 15) = 0.13$; if he wins 70%, with probability $\mathbb{P}_{0.7}(S \geq 15) = 0.42$. Nothing restricts the computation to these three values. The rule’s entire behavior is captured, one parameter value at a time, by the function

$$r(\theta) = \mathbb{P}_{\theta}(S \geq 15), \quad \theta \in [0, 1], \quad (6)$$

its *rejection probability*. Figure 2 draws r over the whole interval; the parameter is unknown but not random, and each point of the curve is a probability computed in the model with θ fixed at that value.

Reading the figure brings back the full form of the dispute, the one section 2.1 set aside for the sake of computation: Karim is *not better*. As sets of parameter values, the hypotheses are

$$H_0 : \theta \leq \tfrac{1}{2} \quad (\Theta_0 = [0, \tfrac{1}{2}]), \quad H_1 : \theta > \tfrac{1}{2} \quad (\Theta_1 = (\tfrac{1}{2}, 1]),$$

the two regions of the figure. The single value $\theta = 1/2$ that every computation so far ran on is the boundary of Θ_0 , and it is no loss: r is increasing in θ (more skill, more wins), so among all

the values the null allows, the boundary is where the rejection probability is largest. (Section 3.3 makes this precise.)

The figure now reads as follows. Over Θ_0 the height of the curve is the probability of a false alarm: rejecting although Karim is not better. Over Θ_1 the same height is the probability of a justified alarm: rejecting because he is. A good rule keeps the first small and the second large: a statistical test controls the probability of rejection wherever the null is true, while trying to make that same probability as large as possible wherever the alternative is true. Every term defined in the rest of this subsection is a reading of r on one of the two regions.

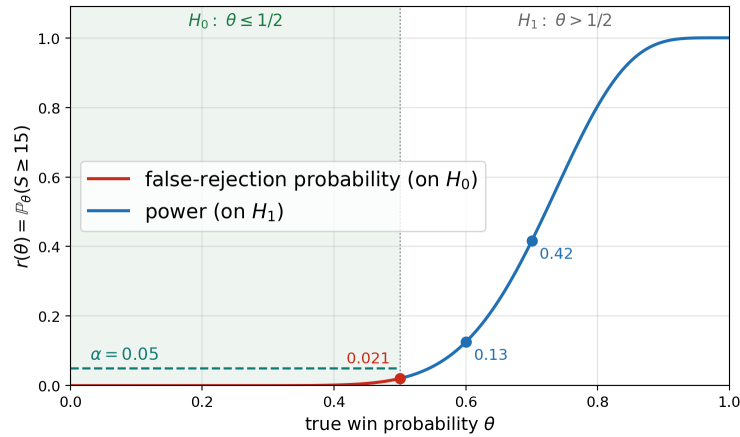


Figure 2: The rejection probability $r(\theta) = \mathbb{P}_\theta(S \geq 15)$ of the rule “reject when $S \geq 15$ ” ($n = 20$), over the whole parameter range. Over the shaded null region its height is a false-rejection probability, largest at the boundary: $r(1/2) = 0.021$, below the α line. Over the alternative region the same height is the power: 0.13 at $\theta = 0.6$, 0.42 at 0.7.

Definition 2.3 (Type I error, size, level)

A *type I error* is the event of rejecting H_0 while the true parameter lies in Θ_0 ; at a given null value θ , it has probability $r(\theta) = \mathbb{P}_\theta(\text{reject } H_0)$. The *size* of a test is its largest false-rejection probability over the null,

$$\text{size} = \sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\text{reject } H_0), \quad (7)$$

and the test *has level* α when its size is at most α .

For the wins rule, the monotonicity of r places the size at the boundary: $\text{size} = r(1/2) = 0.021 \leq 0.05$. The rule has level 5% over the whole null region, and not by luck: the threshold of section 2.4 was a quantile of the boundary null distribution, chosen precisely so that the size stays under α whatever the data turn out to be.

The power. Over the alternative region the heights of r measure what the test is worth, and get their own name:

$$\pi(\theta) = \mathbb{P}_\theta(\text{reject } H_0), \quad \theta \in \Theta_1, \quad (8)$$

the *power function*. Power is a function, not one number: detecting a 51%, a 60% and a 70% player are very different tasks. The wins rule handles them poorly. If Karim truly wins 60% of games, the test detects it with probability 0.13; at a true 70%, probability 0.42; only around 80% does detection become likely at 0.80. Twenty games are simply not much information; detecting a 60% player with power 0.80 takes about 160 games (approximately: for a discrete test the power is not monotone in n).

Figure 3 draws the kind of curve such a rule aims for: still under the α line on all of Θ_0 , then rising fast once θ crosses into Θ_1 . The rise can be steep but never a jump: at fixed n the rejection

probability of the wins rule is a polynomial in θ , hence continuous, and what more data buy is exactly a steeper S-curve (the 160-game figure above is one point of that sharpening). This is the design principle Neyman and Pearson formalized in 1933 [2]: fix the level first, then, among all tests satisfying that level constraint, seek the largest possible power under the alternatives.

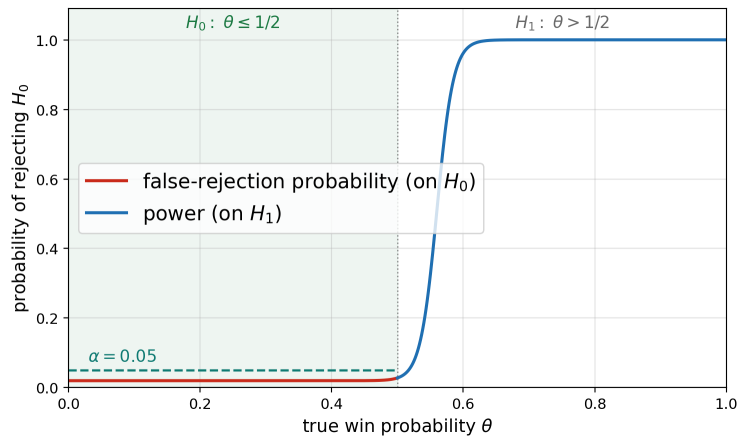


Figure 3: A rejection curve one would wish for, on the same axes: below the α line throughout the null region, close to one throughout most of the alternative. At fixed n the rejection probability is continuous in θ , so the steep rise can never be a jump; larger n steepens it.

The type II error. One failure remains: the test stays silent although Karim really is better. At an alternative value θ , its probability is the complement of the power,

$$\beta(\theta) = \mathbb{P}_{\theta}(\text{do not reject } H_0) = 1 - \pi(\theta), \quad \theta \in \Theta_1. \quad (9)$$

No dial of the construction controls it; it is whatever the level and the sample size leave over. Read off figure 2: $\beta(0.6) = 0.87$ and $\beta(0.7) = 0.58$: at twenty games, a real edge is missed more often than it is caught.

The summary. Crossing what is true with what the test decides:

	H_0 true (Karim not better)	H_1 true (Karim better)
reject H_0	type I error (prob. $r(\theta)$; size $\leq \alpha$)	correct (prob. $\pi(\theta)$)
do not reject	correct	type II error (prob. $\beta(\theta)$)

These numbers give the failure to reject its correct meaning. Karim’s season, $S = 14$ and $p = 0.058$, is compatible with equal skill, *and* compatible with Karim winning 70% of all games, an edge the test misses more often than it catches. Not rejecting H_0 therefore establishes nothing about H_0 : the data were too few, or the effect too small, for the claim to clear the bar. The test measures evidence against the null and never for it; **failure to reject H_0 means insufficient evidence against it, not evidence that it is true**. This is the asymmetry announced in section 2.1, and it is the single most common point of confusion in applied work: *absence of evidence is not evidence of absence*.

Key idea

The testing ladder. Every hypothesis test in this course follows the same basic logic.

1. **Start from the question.** What scientific or practical claim is being assessed? (Is Karim better?)

2. **State H_0 and H_1 .** Translate the question into a null hypothesis to test and an alternative representing the claim of interest.
3. **Choose a statistic.** Reduce the data to a quantity that becomes more extreme when the data depart from H_0 in directions relevant to H_1 .
4. **Determine its behavior under H_0 .** Derive or simulate the distribution of the statistic when the null hypothesis holds.
5. **Interpret the p-value.** If H_0 were true, how probable would a statistic at least as extreme as the one observed be?
6. **Report effect size and uncertainty.** Statistical significance alone does not say how large or important the effect is (section 4).
7. **Check assumptions and design issues.** How sensitive is the conclusion to sample size, distributional shape, dependence, or multiple testing? (Section 5 works out the many-tests case.)
8. **Conclude in context.** Return to the original question: what evidence do the data provide, under the assumptions of the test?

Steps 1–5 have now been walked on the FIFA data; steps 6–8 are the rest of this chapter.

3 The alternative H_1 decides the test

Everything computed so far (the tail, the quantile, the p-value) was read off in the direction of $H_1 : \theta > 1/2$. This section varies the alternative and the null, and once the data, to make one point three times: the alternative, chosen from the question and not from the data, decides what “extreme” means, hence the rejection region, hence the p-value.

3.1 One-sided or two-sided

Yassine, unwilling to concede the framing, restates the dispute: “one of us is better, who knows which.” The claim is now $H_1 : \theta \neq 1/2$, and both directions count as extreme. The p-value collects both tails: on the same 14 wins,

$$p = \mathbb{P}(|S - 10| \geq 4) = 0.115,$$

double the one-sided 0.058, by the symmetry of $\text{Bin}(20, 1/2)$ (figure 4). The same twenty games, summarized by the same statistic, yield 0.058 or 0.115 depending only on the question asked.

Remark

The alternative is chosen before the data. The direction of H_1 must come from the question (is the claim “Karim is better” or “the players differ”?) and must be fixed before the results are seen. Choosing the tail *after* looking (testing $\theta > 1/2$ because Karim happens to lead) silently converts a two-sided question into a one-sided test and doubles the real level: the freedom to pick the favorable tail is worth a factor of two. Section 5 meets the industrial-scale version of this mistake.

3.2 Skewed nulls

The doubling in the previous computation used the symmetry of $\text{Bin}(20, 1/2)$, and symmetry is the exception, not the rule. Karim, on a roll, claims he converts 80% of his penalty kicks. A month of practice data: 16 scored out of 25. The null is now $H_0 : \theta = 0.8$ against $H_1 : \theta \neq 0.8$,

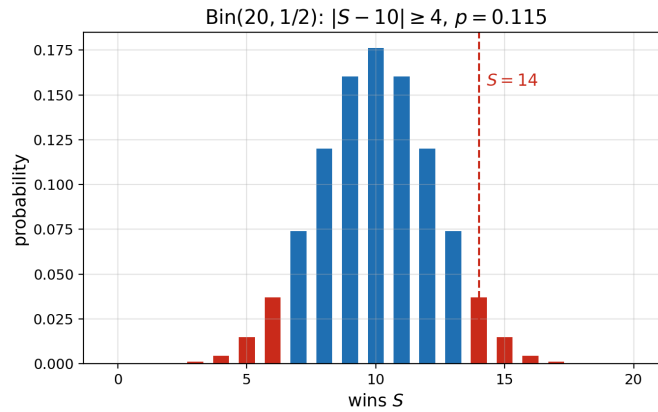


Figure 4: Under the symmetric null $\text{Bin}(20, 1/2)$, the two-sided region $|S - 10| \geq 4$ has two mirror-image tails, and the two-sided p-value doubles the one-sided one.

and the null distribution $\text{Bin}(25, 0.8)$ is visibly skewed (figure 5): the two tails are no longer mirror images, and “as extreme on the other side” has no unique meaning. The observed left tail is $\mathbb{P}(S \leq 16) = 0.047$. Two conventions coexist for the two-sided p-value: *double the smaller tail* ($p = 0.094$), or *sum all outcomes at most as likely as the observed one* ($p = 0.074$ here). They disagree, and neither is wrong; doubling is common in textbooks, while the exact binomial tests of `scipy` and `R` (`binomtest`, `binom.test`) use the second; honest practice states which one is used. The lesson survives beyond binomials: with an asymmetric null distribution, two-sided p-values involve a convention, and the symmetric case of section 3.1 was a coincidence of $\theta_0 = 1/2$.

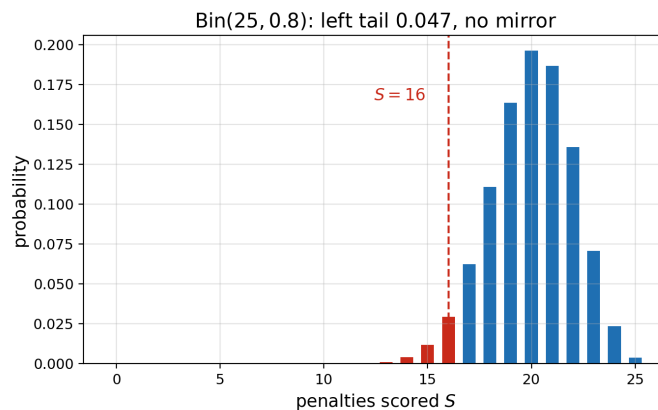


Figure 5: The penalty null $\text{Bin}(25, 0.8)$ is skewed: the observed left tail (0.047) has no mirror image, and the two-sided p-value depends on a convention.

3.3 Composite nulls

Section 2.5 stated the dispute in its full form, $H_0 : \theta \leq 1/2$, a *composite* null: a whole range of parameter values with no single distribution to compute under. It asserted without proof that the rejection probability is largest at the boundary, which is what reduces the composite null to its boundary. Here is the statement.

Property 3.1 (Composite one-sided nulls reduce to the boundary)

For every k , the tail probability $\theta \mapsto \mathbb{P}_\theta(S \geq k)$ is nondecreasing in θ . Consequently

$$\sup_{\theta \leq 1/2} \mathbb{P}_\theta(S \geq k) = \mathbb{P}_{1/2}(S \geq k),$$

so a test of the form “reject when $S \geq k$ ” whose false-rejection probability is at most α at $\theta = 1/2$ has size at most α over the whole composite null $\theta \leq 1/2$.

The monotonicity is intuitive: a better player produces more wins. The smaller θ is, the less often it produces $S \geq k$, so among all the values in the null, $\theta = 1/2$ (the largest, the one closest to the alternative) is the *least favorable* null value, the one where a false rejection is most likely: a false-rejection probability held below α at $\theta = 1/2$ is held below α at every smaller value automatically, so controlling it controls the whole null. Testing $\theta \leq 1/2$ against $\theta > 1/2$ therefore uses the same statistic, the same quantile and the same p-value as testing $\theta = 1/2$ against $\theta > 1/2$. (The proof of the monotonicity is a short coupling argument.) This reduction is general for one-sided problems with this monotonicity, and it is why one-sided tests of a regression coefficient, of the kind met in first year ($\beta \leq 0$ against $\beta > 0$), are computed at the single point $\beta = 0$.

4 Statistical significance is not statistical importance

Karim takes the dispute online and grinds a full ladder season: $n = 2000$ ranked games, a 53% win rate, that is $S = 1060$ wins. The one-sided test of $H_0 : \theta = 1/2$ now rejects emphatically: the binomial tail gives $p = \mathbb{P}(S \geq 1060) = 0.0039$. Highly significant. Has Karim shown he is *much* better? The estimate answers a different question than the test, and both deserve to be reported. The point estimate is $\hat{\theta} = 0.53$, and the 95% confidence interval from the central limit theorem of chapter 1 with a plug-in standard error, $\hat{\theta} \pm 1.96\sqrt{\hat{\theta}(1-\hat{\theta})/n}$, is $[0.508, 0.552]$: an edge of about one to five points (at the midpoint, one extra win every 33 games). The rejection is decisive; the effect is small.

Key idea

Significance is not size. The p-value measures the incompatibility of the data with $\theta = 1/2$, in the direction specified by the alternative; it does not measure the size of the effect. With n large enough, *every* nonzero effect becomes significant: at $n = 2000$, a 53% player rejects fair play with room to spare. Report the estimate and its confidence interval next to any p-value: the p-value alone is a claim about the null value, not about magnitude.

Tests and confidence intervals are, in fact, the same object seen from two sides. The 95% interval $[0.508, 0.552]$ is exactly the set of values θ_0 that the level-5% two-sided Wald test (the statistic $(\hat{\theta} - \theta_0)/\widehat{se}$, with the same estimated standard error) would *not* reject on these data (other tests, such as the exact binomial one, give nearly the same set): $1/2$ falls outside the interval, and the two-sided test of $\theta = 1/2$ rejects; 0.53 falls inside, and the test of $\theta = 0.53$ does not. Reading an interval as “all the nulls these data leave standing” often communicates both significance and size at once, which is why reporting the interval is never a downgrade from reporting the test.

5 Multiple testing

Every number in this chapter was computed under assumptions, most of them loaded as quietly as Arbuthnot’s: independent games, a constant θ and, everywhere, the reading of *one* test at level α . Step 7 of the ladder says to stress-test such assumptions before trusting a verdict. This

section works out the stress test for the last one, the assumption broken most often in practice: what survives of a 5% level when two hundred tests are read at once.

5.1 The smurf screen

The platform hosting the ladder has a smurfing problem: experienced players opening fresh accounts to stomp beginners. The anti-cheat team screens $m = 200$ accounts, testing each one's win rate against $H_0 : \theta = 1/2$ over its last 60 games, at level 5% each. Every individual test is calibrated. The screen as a whole is not. If all 200 accounts are honest, the expected number of false flags is at most $200 \times 0.05 = 10$ (about 9 for this discrete test, which rejects at $S \geq 37$ with real size 0.046), and for independent tests the probability of at least one false flag is

$$\mathbb{P}(\text{at least one false rejection}) = 1 - (1 - \alpha)^m, \quad (10)$$

already 40% at $m = 10$ and 0.99996 at $m = 200$ (figure 6). The quantity on the left is called the *family-wise error rate* (FWER). A per-test level does not control it: with enough calibrated tests, at least one false discovery becomes almost certain. This is the large-scale version of the tail-picking of section 3.1: many looks at noise, each calibrated on its own, together very likely to flag something.

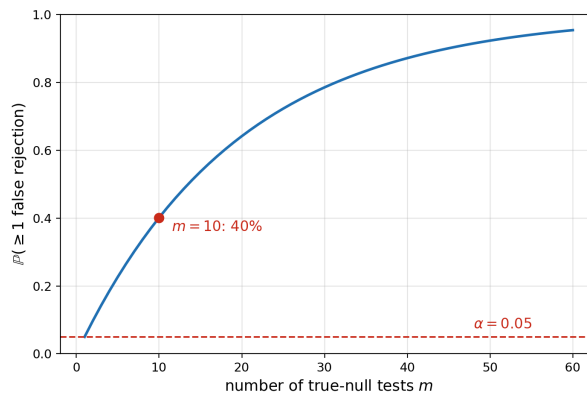


Figure 6: The probability (10) of at least one false rejection among m independent true-null tests at $\alpha = 5\%$: 40% at $m = 10$, near certainty long before $m = 200$.

5.2 Controlling the family-wise error

The oldest fix is *Bonferroni*: test each account at level α/m instead of α . By the union bound, which requires no independence, the probability that *any* true null is rejected is at most $m \cdot \alpha/m = \alpha$. The price is severity: at $m = 200$ the working threshold is $0.05/200 = 0.00025$, and moderate effects are lost. *Holm's procedure* spends the same budget more efficiently (sort the p-values, compare the smallest to α/m , the next to $\alpha/(m-1)$, and so on until the first failure) and rejects everything Bonferroni rejects, sometimes more, with the same FWER guarantee (stated here without proof).

5.3 Controlling the false discovery rate

For a screening problem like the smurf hunt, guarding against any single false flag is often too strict; what matters is that the flagged *list* be mostly right. The *false discovery rate* (FDR) is the expected proportion of false flags among the flags, and the *Benjamini-Hochberg* (BH) procedure controls it at level α [3]: sort the p-values $p_{(1)} \leq \dots \leq p_{(m)}$, find the largest k such that

$$p_{(k)} \leq \frac{k}{m} \alpha, \quad (11)$$

and flag the k smallest. The threshold is a rising line rather than a floor: the more genuinely small p-values the screen contains, the more the procedure lets through. The BH theorem covers independent tests and extends to certain positively dependent families, not to arbitrary dependence; Bonferroni and Holm, resting on the union bound alone, need no such qualification.

The simulated screen makes the comparison concrete. Among the 200 accounts, 195 are honest and 5 are planted smurfs with a true 80% win rate. Naive per-account testing at 5% flags 20 accounts: the 5 smurfs and 15 innocent players (on the high side, since about 9 false flags are expected among 195 honest accounts, and 15 or more occur with probability about 0.04). Bonferroni flags exactly the 5 smurfs, and so does BH (figure 7); with weaker smurfs BH would typically keep more of the real ones than Bonferroni, at the cost of a controlled fraction of false flags.

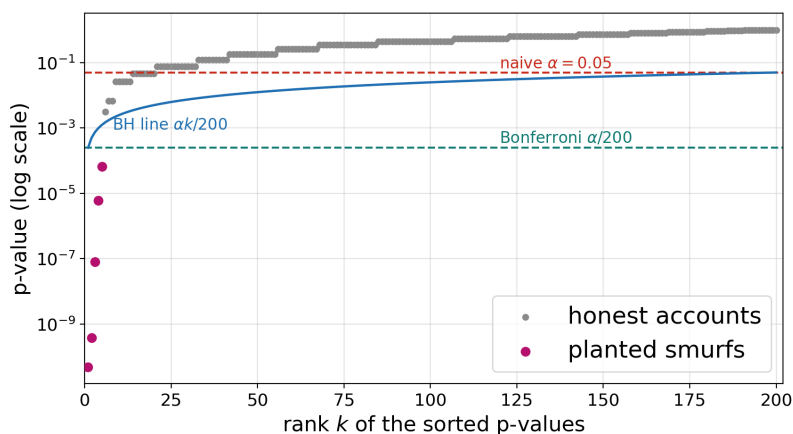


Figure 7: The 200 sorted p-values of the smurf screen (log scale). The five planted smurfs sit far below the Bonferroni threshold $\alpha/200$; fifteen honest accounts fall between the BH line and the naive 5% line and are flagged only by the naive screen. The staircase pattern in the honest p-values is the discreteness of section 2.4 again: with 60 games, only 61 p-values are possible.

Reality check

Which correction to use is a question about the purpose of the analysis, because the two guarantees differ in strictness. FWER control is the conservative one: it bounds the probability of even a single false flag anywhere in the family, and pays for that with fewer discoveries. FDR control is more permissive: it flags more, and guarantees only that the flagged list is, on average, no more than a fraction α false flags. A confirmatory analysis, whose conclusion will support a published claim or a costly decision, controls the FWER. An exploratory screen over many candidates controls the FDR, accepting that expected fraction of false leads among its discoveries in exchange for more of them. In both cases the correction is chosen *before* the p-values are examined, for the same reason the level and the alternative are: a procedure adjusted after seeing what it saves no longer controls any error rate.

6 Conclusion

The chapter’s product is the testing ladder boxed in section 2.5, now climbed in full:

step	in this chapter
1. the question	is Karim better, or lucky?
2. H_0 vs H_1	$\theta \leq 1/2$ against $\theta > 1/2$; the alternative decided every tail
3. the statistic	the number of wins S
4. null distribution	$\text{Bin}(20, 1/2)$, exactly or by simulation
5. the p-value	0.058: a gray zone, not a verdict
6. effect size	2000 games: $p = 0.0039$ yet $\hat{\theta} \in [0.508, 0.552]$: significant and small
7. stress tests	200 honest accounts at 5%: false flags almost certain; Bonferroni/Holm or BH repair
8. conclude in context	evidence under assumptions, never an automatic verdict

Failing to reject the null means insufficient evidence against it, never evidence that it is true: twenty games could not distinguish equal skill from a 70% edge. The alternative, chosen from the question before the data, decides the rejection region, the quantile, and the p-value. Significance is not size: enough games make a trivial edge arbitrarily significant. And every level ever quoted rests on assumptions (independence, distributional shape, a single test) that can fail as silently as Arbuthnot’s fair coin was assumed; his overreach, reading a cause into a rejection, is the standing warning of step 8.

The next chapter puts this ladder to work on a real model: a regression of customer spending on its covariates, with a t -test for every coefficient and a screen of twenty-eight coefficients at once, the multiple-testing problem of section 5 in a regression.

References

- [1] J. Arbuthnot. An argument for Divine Providence, taken from the constant regularity observ’d in the births of both sexes. *Philosophical Transactions of the Royal Society*, 27:186–190, 1710.
- [2] J. Neyman and E. S. Pearson. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A*, 231:289–337, 1933.
- [3] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57(1):289–300, 1995.