***Learning Statistics with R - University of Adelaide***

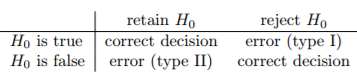
***Part IV – Statistical Theory***

**Chapter 11 – Hypothesis Testing**

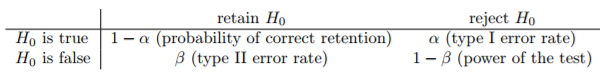
* “The process of induction is the process of assuming the simplest law that can be made to harmonize w/ our experience. This process, however, has no logical foundation but only a psychological one. It is clear there are no grounds for believing that the simplest course of events will really happen. It is a hypothesis that the sun will rise tomorrow: and this means that we do not know whether it will rise.” – Ludwig Wittgenstein
* **Estimation** was 1 of the 2 big ideas in **inferential statistics** + the other big idea is **hypothesis testing**.
* In its most abstract form, hypothesis testing really a very simple idea: researcher has some theory about the world + wants to determine whether or not the data actually support that theory.
* However, details are messy + most people find the theory of hypothesis testing to be the most frustrating part of statistics.
* Simple example study: Seek to test whether clairvoyance exists 🡺 Each participant sits at a table + is shown a card by an experimenter, which is black on 1 side + white on the other.
* Experimenter takes card away + places it on a table in an adjacent room black or white side up completely at random, w/ the randomization occurring only after experimenter has left the room w/ the participant.
* 2nd experimenter comes in + asks the participant which side of the card is facing upwards.
* Purely a 1-shot experiment: Each person sees only 1 card + gives only 1 answer + *at no stage is the participant actually in contact w/ someone who knows the right answer.*
* Dataset is very simple = asked the question of N = 100 people + some number X = 62 got the answer right, a surprisingly large number, sure, but is it large enough to claim evidence for ESP?
* This is the situation where hypothesis testing comes in useful.
* 1st distinction you need to keep clear is between **research hypotheses** and **statistical hypotheses**.
* In ESP study, overall scientific goal = to demonstrate clairvoyance exists
* Clear research goal: hoping to discover *evidence* for ESP
* In other situations, might actually be more neutral than that, so might say research goal = to determine whether or not clairvoyance *exists*.
* Basic point: a **research hypothesis** involves making a substantive, testable scientific claim
* If you’re a psychologist, your research hypotheses are fundamentally about psychological constructs
* Any of the following would count as research hypotheses:
* Listening to music reduces ability to pay attention to other things = a claim about causal relationship between 2 psychologically meaningful concepts (listening to music + paying attention to things), so it’s a perfectly reasonable research hypothesis.
* Intelligence is related to personality = a relational claim about 2 psychological constructs (intelligence + personality), but claim is weaker: *correlational*, NOT causal.
* Intelligence is speed of information processing: This hypothesis has a quite different character:
* not actually a relational claim at all but an **ontological** claim about the *fundamental character of intelligence*
* Actually worth expanding on this one
* Usually easier to think about how to construct experiments to test research hypotheses of the form “*does X affect Y?”* than to address claims like “*what is X?*”
* In practice, usually you find ways of *testing relational claims that follow from ontological ones.*
* Ex: If I believe intelligence is speed of information processing in the brain, my experiments will often involve looking for *relationships* between *measures of intelligence* + *measures of* *speed*.
* As a consequence, most everyday research questions tend to be **relational** in nature, but are almost always motivated by *deeper ontological questions* about the *state of nature*.
* Notice in practice, research hypotheses could overlap a lot.
* Ultimate goal in ESP experiment might be to test an ontological claim “ESP exists”, but I might operationally restrict myself to a *narrower* hypothesis like “Some people can ‘see’ objects in a clairvoyant fashion”.
* That said, there are some things that really don’t count as proper research hypotheses in any meaningful sense:
* Love is a battlefield. too vague to be testable.
* While it’s okay for a research hypothesis to have a degree of vagueness to it, it has to be possible to **operationalize** theoretical ideas
* Difficult to see how this can be converted into any concrete research design.
* If that’s true, this isn’t a scientific research hypothesis, it’s a pop song.
* Doesn’t mean it’s not interesting: a lot of deep questions humans have fall into this category
* Maybe 1 day science will be able to construct testable theories of love, or to test to see if God exists, + so on; but right now we can’t
* The first rule of tautology club is the first rule of tautology club: Not a substantive claim of any kind
* True *by definition*: No conceivable state of nature could possibly be inconsistent w/ this claim
* As such, say this = an **unfalsifiable hypothesis** + as such it is outside the domain of science
* *Whatever else you do in science, claims must have the possibility of being wrong.*
* More people in my experiment will say “yes” than “no”: Fails as a research hypothesis b/c it’s a claim *about the data set*, not *about the psychology* (unless your actual research question is whether people have some kind of “yes” bias).
* This hypothesis is starting to sound more like a **statistical hypothesis** than research hypothesis
* **Research Hypotheses** can be somewhat messy at times + ultimately they *are* scientific claims.
* **Statistical hypotheses** are *neither* of these 2 things 🡺 MUST be mathematically precise + MUST correspond to specific claims about the *characteristics of the data generating mechanism* (i.e., the “population”).
* Even so, the intent is that statistical hypotheses bear a *clear relationship* to the substantive research hypotheses you care about
* Ex: ESP study 🡪 research hypothesis = some people are able to see through walls/whatever.
* What I want to do is to *map* this onto a statement about *how* data were generated
* Quantity I’m interested in w/in the experiment is P(correct), the true-*but-unknown* probability w/ which participants in my experiment answer the question correctly.
* Let’s use the Greek letter θ (theta) to refer to this probability.
* Here are 4 different statistical hypotheses:
* If ESP doesn’t exist + if my experiment is well designed, my participants are just guessing:
* should expect them to get it right 1/2 of the time
* so my statistical hypothesis is the true probability of choosing correctly is θ = 0.5.
* Suppose ESP does exist + participants can see the card.
* If true, people will perform better *than chance*.
* Statistical hypothesis is θ > 0.5.
* ESP does exist, but the colors are all reversed + people don’t realize it
* If that’s how it works you’d expect people’s performance to be below *chance*.
* Correspond to a statistical hypothesis that θ < 0.5.
* Suppose ESP exists, but I have no idea whether people are seeing the right or wrong color.
* Only claim I could make about the data would be the probability of making the correct answer is not equal to 50%
* Corresponds to the statistical hypothesis that θ != 0.5.
* All of these are legitimate examples of a statistical hypothesis b/c they are statements about a population parameter + are meaningfully related to my experiment.
* What this discussion hopefully makes clear is that *when attempting to construct a statistical hypothesis, test that the researcher actually has 2 quite distinct hypotheses to consider.*
* 1st: They have a **research hypothesis** (claim about psychology)
* 2nd: It corresponds to a **statistical hypothesis** (claim about the data-generating population).
* ESP example: these might be:
* Research hypothesis: “ESP exists”
* Statistical Hypothesis: θ != 0.5
* The key thing to recognize is a statistical hypothesis test is a test of the statistical hypothesis, NOT the research hypothesis.
* If a study is badly designed, the link between the research + statistical hypothesis is broken.
* Suppose the ESP study was conducted in a situation where participants can actually see the card reflected in a window
* if that happens, I’d be able to find very strong evidence that θ != 0.5, but this would tell us nothing about whether “ESP exists”.
* So, I have a **research hypothesis** that *corresponds to what I want to believe about the world* + can map it onto a **statistical hypothesis** that *corresponds to what I want to believe about how the data were generated*.
* **Null** **hypothesis, H0**, corresponds to *the exact opposite of what I want to believe*
* Now turn to focus exclusively on that, almost to the neglect of the thing I’m actually interested in, **alternative hypothesis, H1**
* ESP example 🡺 null = θ = 05, since that’s what we’d expect if ESP didn’t exist.
* Hope is that ESP is real + the alternative to this null is θ != 0.5.
* Dividing up the possible values of θ into 2 groups: those values I hope aren’t true (null) + those I’d be happy w/ if they turn out to be right (alternative). Having
* Important thing = *Recognize that the goal of a hypothesis test is NOT to show the alternative hypothesis is (probably) true but to show that the null hypothesis is (probably) false.*
* Ex: Hypothesis test = a criminal trial of the null hypothesis (defendant), researcher = prosecutor, + statistical test itself = judge.
* There is a *presumption of innocence*: null hypothesis is deemed to be true unless the researcher can prove *beyond a reasonable doubt* it is false.
* Free to design the experiment in any way (w/in reason) w/ goal to maximize the chance the data will yield a conviction for the crime of being false.
* The catch = Statistical test sets rules of a trial + they are *designed to protect null* 🡺 specifically to ensure if the null is actually true, the chances of a false conviction *are guaranteed to be low*

**11.2 Two types of errors**

* Want to construct test so we never make any errors 🡺 never possible.
* Always have to accept there’s a chance we did the wrong thing + as a consequence, the goal behind statistical hypothesis testing is NOT to eliminate errors, but to *minimize them*.
* It is either the case the null is true or it is false + our test will either **reject the null** or **retain it**
* After we run the test + make our choice, 1 of 4 things might have happened:

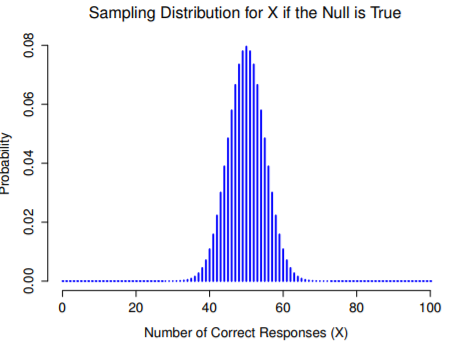


* **Type I Error** = reject a null that is actually true
* **Type II Error** = retain a null when it is in fact false
* A criminal trial requires you establish “beyond a reasonable doubt” that the defendant did it.
* All evidentiary rules are (in theory, at least) designed to ensure there’s (almost) no chance of wrongfully convicting an innocent defendant (trial is designed to protect the rights of a defendant)
* In other words, a criminal trial doesn’t treat the 2 types of error in the same way
* Punishing the innocent is deemed to be *much worse* than letting the guilty go free.
* Statistical tests are pretty much the same
* The Single Most Important Design Principle Of The Test = to *control the probability of a type I error + keep it below some fixed probability*.
* This probability, α, = the **significance level** of the test (sometimes the **size** of the test).
* A hypothesis test is said to have significance level α if the type I error rate is no larger than α.
* Would also like to keep type II error rate under control too, denoted w/ β.
* Much more common to refer to the **power** of a test = probability w/ which we reject a null when it *really is false* (good) 🡺 **1 – β**



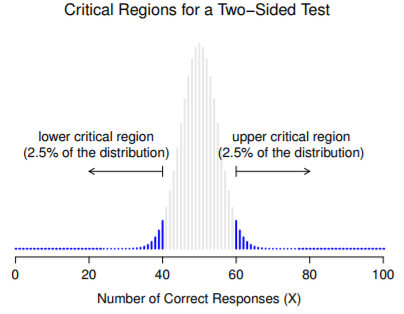
* A powerful hypothesis test = one that has a *small value of β* while *still keeping α fixed at some (small) desired level*.
* By convention, scientists make use of 3 different α levels: .05, .01, .001.
* Tests are designed to ensure the α level is kept small, but *there’s no corresponding guarantee regarding β*.
* Certainly would like type II error rate to be small + we try to design tests that keep it small, but this is very much secondary to the overwhelming need to control the type I error rate.
* It is better to retain 10 false nulls than to reject a single true one”.
* 1 thing to avoid = the word “prove”
* a statistical test really doesn’t *prove* a hypothesis is true or false.
* Proof implies certainty + statistics = never having to say you’re certain.
* Some argue you’re only allowed to make statements like “rejected the null”, “failed to reject the null”, or possibly “retained the null”

**11.3 Test statistics and sampling distributions**

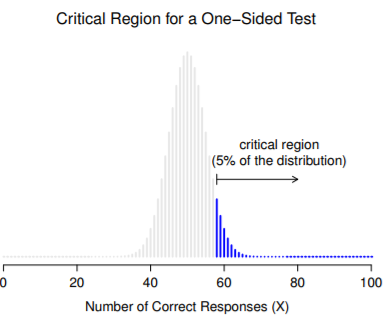
* ESP 🡺 ignore actual data obtained for the moment + think about the structure of the experiment
* The form of the data is that “X out of N people correctly IDed the color of the hidden card”
* Moreover, suppose the null really is true: ESP doesn’t exist, + the true probability anyone picks the correct color is exactly θ = 0.5.
* Would then expect the proportion of people w/ a correct response to be pretty close to 50%
* We’d say “X/N is approximately 0.5” (don’t expect this fraction to be exactly 0.5)
* If X = 99 participants got the question right, we’d feel pretty confident the null is wrong.
* if only X = 3 got the answer right, similarly confident the null was wrong.
* We have a quantity X we can calculate by looking at data + after looking at X, we make a decision about whether to believe the null is correct, or to reject it in favor of the alternative.
* **Test statistic** = what we calculate to guide our choices.
* Having chosen a test statistic, state precisely *which values* of the test statistic cause us to reject the null + which values cause us to keep it.
* In order to do so, *need to determine what the sampling distribution of the test statistic would be if the null were actually true*
* This distribution tells us *exactly what values of X our null would lead us to expect*, + therefore, we can use this it as a tool for assessing how closely the null agrees w/ our data.
* To determine the sampling distribution of a test statistic for a lot of hypothesis tests = complicated
* Sometimes it’s very easy + fortunately, ESP example provides 1 of the easiest cases.
* Our population parameter θ = just the overall probability people respond correctly when asked the question, + our test statistic X = count of people who did so out of a sample size of N.
* That’s exactly what the binomial distribution describes
* We’d say “the null predicts X is binomially distributed” = X ~ Binomial(θ, N)
* Since the null states θ = 0.5 + our experiment has N = 100, we have the sampling distribution needed
* 
* No surprises really: the null says X = 50 is the most likely outcome + says we’re almost certain to see somewhere between 40-60 correct responses.

**11.4 Making decisions**

* We’ve constructed a test statistic (X) + we chose this test statistic in such a way that we’re pretty confident if X is close to N/2, we should retain the null, + if not we should reject it.
* But exactly *which* values of the test statistic do associate w/ the null + which w/ the alternative?
* ESP: observed a value of X = 62. What to make? Believe the null or the alternative?
* The **critical region** of the test corresponds to values of X that lead to rejecting the null
* Consider what we know:
* X should be very big or very small in order to reject the null
* If the null is true, the sampling distribution of X is Binomial(0.5, N)
* If α = .05, the critical region must cover 5% of this sampling distribution
* It’s important to understand the **critical region** corresponds to values of X for which we’d reject the null + the *sampling distribution* in question *describes the probability* we’d obtain a *particular value of X* if the null were actually true.
* Suppose we chose a critical region that covers 20% of the sampling distribution + that the null is actually true 🡺 The probability of incorrectly rejecting the null = 20% 🡺 20% of getting our test statistic if the null were true
* Therefore, we’d have built a test that had α = 0.2.
* If we want α = .05, the critical region is only allowed to cover 5% of the sampling distribution of our test statistic.



* The critical regions associated w/ the hypothesis test for ESP w/ significance level α = 0.05.
* The plot shows the sampling distribution of X under the null hypothesis + the grey bars correspond to values of X for which we’d *retain the null*.
* Black bars = start of the critical regions = values of X for which we’d reject the null.
* B/c the alternative is 2-sided (allows both θ < 0.5 + θ > 0.5), the critical region covers *both tails* of the distribution.
* To ensure an α = 0.05, must ensure each of the 2 regions encompasses 2.5% of the sampling distribution
* As it turns out, those 3 things uniquely solve the problem:
* Our critical region = the most extreme values (**tails**) of the distribution.
* If we want α = 0.05, our critical regions correspond to X < 40 + X > 60
* If the # of people saying “true” is between 41-59, we should retain the null.
* If the # is between 0-40 or 60-100, we should reject the null.
* 40 + 60 are referred to as the **critical values**, since they define the edges of the critical region.
* Strictly speaking, the test just constructed has α = 0.057, which is a bit too generous.
* If I’d chosen 39 + 61 to be the boundaries for the critical region, the critical region only covers 3.5% of the distribution.
* It makes more sense to use 40 + 60 as my critical values + be willing to tolerate a 5.7% type I error rate, since that’s as close as we can get to a value of α = 0.05
* At this point, our hypothesis test is essentially complete
* (1) We choose an α level (e.g., α = 0.05
* (2) Came up w/ some test statistic (X) that does a good job (in some meaningful sense) of comparing H0 to H1
* (3) Figured out the sampling distribution of the test statistic on the assumption the null is true (binomial)
* (4) Calculated the critical region that produces an appropriate α level (0-40 + 60-100).
* All we have to do now is calculate the value of the test statistic for the real data, X = 62, + then compare it to the critical values to make our decision.
* Since 62 is greater than the critical value 60, we reject the null + say the test has produced a **significant result**
* The concept of **statistical significance** is actually a very simple one but has a very unfortunate name.
* If the data allow us to reject the null, we say “the result is statistically significant”, which is often shortened to “the result is significant”.
* **Significant** just means something like “indicated”, rather than “important”.
* **Statistically Significant** means is the data allowed us to reject a null
* Whether or not the result is *actually important* in the real world is a very different question, + depends on all sorts of other things
* If we take a moment to think about the statistical hypotheses so far, H0 : θ = 0.5 and H1 : θ != 0.5 , we notice the alternative covers both the possibility that θ > 0.5 + the possibility that θ < 0.5.
* This makes sense if we really think ESP could produce better-than-chance performance or worse-than-chance performance (there are some people who think that).
* This is an example of a **2-sided test** b/cthe alternative covers the area on both sides of the null + as a consequence, the critical region of the test covers both tails of the sampling distribution (2.5% on either side if α = 0.05)
* It might be the case I’m only willing to believe in ESP if it produces *better than chance* performance.
* If so, my alternative only covers the possibility that θ > 0.5 + as a consequence the null now becomes θ <= 0.5 and therefore H0 : θ <= 0.5 and H1 : θ > 0.5
* This is a **one-sided test** + when this happens the critical region only covers 1 tail of the sampling distribution



* The critical region for a 1-sided test when the alternative is θ = 0.05
* We only reject the null for large values of X.
* As a consequence, the critical region only covers the upper tail of the sampling distribution, specifically the upper 5% of the distribution

**11.5 The p value of a test**

* In one sense, our hypothesis test is complete: we’ve constructed a test statistic, figured out its sampling distribution if the null is true, + constructed the critical region for the test.
* Nevertheless, I’ve actually omitted the most important number of all: the p value. It is to this topic that we now turn. There are two somewhat different ways of interpreting a p value, one proposed by Sir Ronald Fisher and the other by Jerzy Neyman. Both versions are legitimate, though they reflect very different ways of thinking about hypothesis tests. Most introductory textbooks tend to give Fisher’s version only, but I think that’s a bit of a shame. To my mind, Neyman’s version is cleaner, and actually better reflects the logic of the null hypothesis test. You might disagree though, so I’ve included both. I’ll start with Neyman’s version... 11.5.1 A softer view of decision making One problem with the hypothesis testing procedure that I’ve described is that it makes no distinction at all between a result this “barely significant” and those that are “highly significant”. For instance, in my ESP study the data I obtained only just fell inside the critical region - so I did get a significant effect, but was a pretty near thing. In contrast, suppose that I’d run a study in which X “ 97 out of 7The internet seems fairly convinced that Ashley said this, though I can’t for the life of me find anyone willing to give a source for the claim. - 336 - my N “ 100 participants got the answer right. This would obviously be significant too, but my a much larger margin; there’s really no ambiguity about this at all. The procedure that I described makes no distinction between the two. If I adopt the standard convention of allowing α “ .05 as my acceptable Type I error rate, then both of these are significant results. This is where the p value comes in handy. To understand how it works, let’s suppose that we ran lots of hypothesis tests on the same data set: but with a different value of α in each case. When we do that for my original ESP data, what we’d get is something like this Value of α 0.05 0.04 0.03 0.02 0.01 Reject the null? Yes Yes Yes No No When we test ESP data (X “ 62 successes out of N “ 100 observations) using α levels of .03 and above, we’d always find ourselves rejecting the null hypothesis. For α levels of .02 and below, we always end up retaining the null hypothesis. Therefore, somewhere between .02 and .03 there must be a smallest value of α that would allow us to reject the null hypothesis for this data. This is the p value; as it turns out the ESP data has p “ .021. In short: p is defined to be the smallest Type I error rate (α) that you have to be willing to tolerate if you want to reject the null hypothesis. If it turns out that p describes an error rate that you find intolerable, then you must retain the null. If you’re comfortable with an error rate equal to p, then it’s okay to reject the null hypothesis in favour of your preferred alternative. In effect, p is a summary of all the possible hypothesis tests that you could have run, taken across all possible α values. And as a consequence it has the effect of “softening” our decision process. For those tests in which p ď α you would have rejected the null hypothesis, whereas for those tests in which p ą α you would have retained the null. In my ESP study I obtained X “ 62, and as a consequence I’ve ended up with p “ .021. So the error rate I have to tolerate is 2.1%. In contrast, suppose my experiment had yielded X “ 97. What happens to my p value now? This time it’s shrunk to p “ 1.36 ˆ 10´25, which is a tiny, tiny8 Type I error rate. For this second case I would be able to reject the null hypothesis with a lot more confidence, because I only have to be “willing” to tolerate a type I error rate of about 1 in 10 trillion trillion in order to justify my decision to reject. 11.5.2 The probability of extreme data The second definition of the p-value comes from Sir Ronald Fisher, and it’s actually this one that you tend to see in most introductory statistics textbooks. Notice how, when I constructed the critical region, it corresponded to the tails (i.e., extreme values) of the sampling distribution? That’s not a coincidence: almost all “good” tests have this characteristic (good in the sense of minimising our type II error rate, β). The reason for that is that a good critical region almost always corresponds to those values of the test statistic that are least likely to be observed if the null hypothesis is true. If this rule is true, then we can define the p-value as the probability that we would have observed a test statistic that is at least as extreme as the one we actually did get. In other words, if the data are extremely implausible according to the null hypothesis, then the null hypothesis is probably wrong. 11.5.3 A common mistake Okay, so you can see that there are two rather different but legitimate ways to interpret the p value, 8That’s p “ .000000000000000000000000136 for folks that don’t like scientific notation! - 337 - one based on Neyman’s approach to hypothesis testing and the other based on Fisher’s. Unfortunately, there is a third explanation that people sometimes give, especially when they’re first learning statistics, and it is absolutely and completely wrong. This mistaken approach is to refer to the p value as “the probability that the null hypothesis is true”. It’s an intuitively appealing way to think, but it’s wrong in two key respects: (1) null hypothesis testing is a frequentist tool, and the frequentist approach to probability does not allow you to assign probabilities to the null hypothesis... according to this view of probability, the null hypothesis is either true or it is not; it cannot have a “5% chance” of being true. (2) even within the Bayesian approach, which does let you assign probabilities to hypotheses, the p value would not correspond to the probability that the null is true; this interpretation is entirely inconsistent with the mathematics of how the p value is calculated. Put bluntly, despite the intuitive appeal of thinking this way, there is no justification for interpreting a p value this way. Never do it.