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

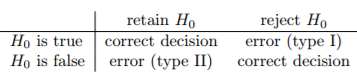
***Part IV – Statistical Theory***

**Chapter 11 – Hypothesis Testing**

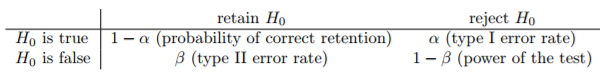
* “The process of induction is the process of assuming the simplest law that can be made to harmonize w/ our experience. This process, however, has no logical foundation but only a psychological one. It is clear there are no grounds for believing that the simplest course of events will really happen. 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:

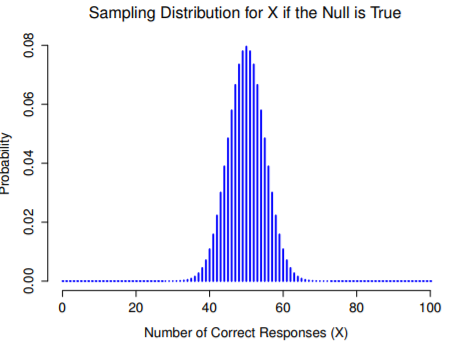


* **Type I Error** = reject a null that is actually true
* **Type II Error** = retain a null when it is in fact false
* A criminal trial requires you establish “beyond a reasonable doubt” that the defendant did it.
* All evidentiary rules are (in theory, at least) designed to ensure there’s (almost) 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**



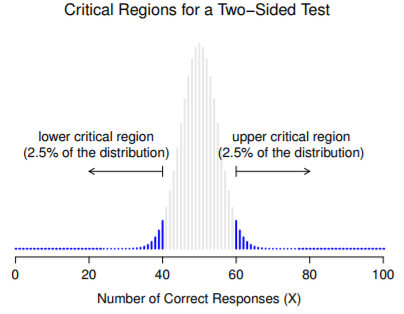
* A powerful hypothesis test = one that has a *small value of β* while *still keeping α fixed at some (small) desired level*.
* By convention, scientists make use of 3 different α levels: .05, .01, .001.
* Tests are designed to ensure the α level is kept small, but *there’s no corresponding guarantee 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**

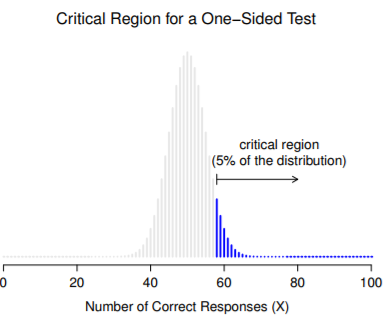
* ESP 🡺 ignore actual data obtained for the moment + think about the structure of the experiment
* The form of the data is that “X out of N people correctly IDed the color of the hidden card”
* Moreover, suppose the null really is true: ESP doesn’t exist, + the true probability anyone picks the correct 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**

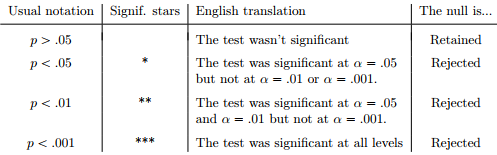
* In one sense, our hypothesis test is complete: we’ve constructed a test statistic, figured out its sampling distribution if the null is true, + constructed the critical region for the test.
* 2 somewhat different ways of interpreting a **p value**, 1 proposed by Sir Ronald Fisher + 1 by Jerzy Neyman.
* Both are legitimate, though reflect very different ways of thinking about hypothesis tests.
* Most introductory texts give Fisher’s version only, but one could say Neyman’s version is cleaner + actually better reflects the logic of the null hypothesis test.
* 1 problem w/ the hypothesis testing procedure described is it makes no distinction at all between a *barely* + *highly* significant.
* Suppose we ran lots of hypothesis tests on the same data set but w/ a different α in each.



* Using α levels >= .03, we always reject the null
* Somewhere between .02-.03 there must be a smallest value of α that allows us to reject the null for this data; This is the **p value**
* **p** = the *smallest Type I error rate (α)* you have to be willing to *tolerate* if you want to reject the null
* If p describes an error rate you find intolerable, you must retain the null.
* If comfortable w/ an error rate = p, it’s okay to reject the null in favor of the preferred alternative.
* In effect, **p** = a summary of all possible hypothesis tests you could have run, taken across all possible α values.
* As a consequence, it has the effect of **softening** our decision process.
* For those tests in which p <= α you’d have rejected the null, whereas for those tests in which p > α you’d have retained the null.
* ESP study: X = 62 w/ p = .021 🡪 the error rate I have to tolerate is 2.1%.
* Suppose experiment yielded X = 97 🡪 p shrinks to 1.36\*10^-25, a tiny, tiny Type I error rate.
* For this 2nd case I’d be able to reject the null w/ a lot more confidence, b/c I only have to be *willing to tolerate* a type I error rate of about 1 in 10 trillion in order to justify my decision to reject.
* The 2nd definition of a p-value comes from Sir Ronald Fisher (in most introductory stats textbooks)
* Notice how when we constructed the critical region, it corresponded to the tails (extreme values) of a sampling distribution?
* That’s not a coincidence: almost all “good” tests have this characteristic (good in the sense of minimizing type II error rate, β).
* The reason for that is a *good critical region almost always corresponds to those values of the test statistic least likely to be observed if the null is true.*
* If this is true, we can define p-value = the probability we observe 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, the null is probably wrong
* 2 rather different but legitimate ways to interpret the p value, but a mistaken approach is to refer to the p value as “the probability the null is true”.
* Wrong in 2 key respects
* (1) null hypothesis testing is a frequentist tool + the frequentist approach to probability does not allow you to assign probabilities to the null
* According to this view of probability, the null is either true or it is not; it cannot have a “5% chance” of being true.
* (2) Even w/in the Bayesian approach, which does let you assign probabilities to hypotheses, the p value wouldn’t correspond to the probability that the null is true
* This interpretation is entirely inconsistent w/ the mathematics of how the p value is calculated.

**11.6 Reporting the results of a hypothesis test**

* When writing up the results of a hypothesis test, there’s usually several pieces of info you need to report, but it varies a fair bit from test to test.
* Regardless of what test you’re doing, the 1 thing you always have to do is say something about the p-value + whether or not the outcome was significant.
* The fact you have to do this is unsurprising as it’s the whole point of doing the test.
* What might be surprising is the fact there’s some contention over exactly how to do it.
* A certain amount of tension exists regarding whether or not to report the exact p-value obtained, or to state only that p > α for a significance level chosen in advance
* To see why this is an issue, the key thing to recognize is p-values are terribly convenient.
* In practice, the fact we can compute a p-value means we don’t actually have to specify any α level at all in order to run a test.
* Can calculate a p-value + interpret it directly 🡪 if you get p = .062, it means you’d have to be willing to tolerate a Type I error rate of 6.2% to justify rejecting the null + if you personally find 6.2% intolerable, you retain the null.
* Why don’t we just report the actual p-value + let the reader make up their own minds about what an acceptable Type I error rate is?
* This approach has a big advantage of softening the decision-making process
* If you accept the *Neyman* definition of a p-value, that’s the whole point of the p value: We no longer have a fixed significance level of α = .05 as a bright line separating “accept” from “reject” decisions + it removes the rather pathological problem of being forced to treat p = .051 in a fundamentally different way to p = .049.
* This flexibility is both the advantage + disadvantage to the p value.
* A lot of people don’t like reporting an exact-p value is it gives the researcher a bit too much freedom + it lets you change your mind about what error tolerance you’re willing to put up w/ after you look at the data.
* Ex: Suppose I ran a test + ended up w/ p = .09. Should I accept or reject?
* Now that I’ve looked at the data, I’m starting to think a 9% error rate isn’t so bad, especially compared to how annoying it’d be to have to admit to my experiment failed.
* So, to avoid looking like I just made it up after the fact, I now say that my α = .1: a 10% type I error rate isn’t too bad, + at this level, my test is significant
* The worry here is that temptation to shade things a little bit here + there is really, really strong.
* As anyone who has ever run an experiment can attest, it’s a long + difficult process, + you often get very attached to hypotheses + it’s hard to let go + admit an experiment didn’t find what you wanted it to find.
* If we use the “raw” p-value, people will start interpreting data in terms of *what they want to believe*, NOT what the data are actually saying
* In this view, one must specify α in advance + then only report whether the test was significant or not to keep ourselves honest.
* Rare for a researcher to specify a single α level ahead of time in practice + convention = rely on 3 standard significance levels: .05, .01 and .001.
* When reporting results, indicate which (if any) of these significance levels allow you to reject the null 🡪 allows us to soften the decision rule a little bit, since p < .01 implies that the data meet a stronger evidentiary standard than p < .05 would.



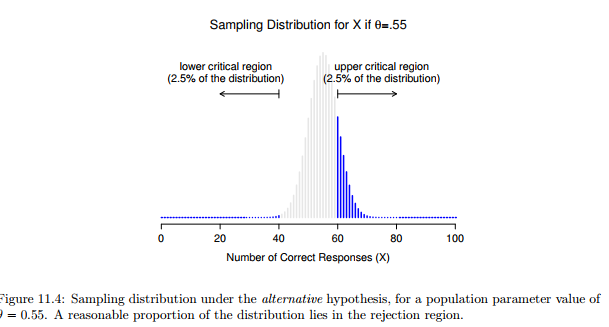
* Since these levels are fixed in advance by convention, it prevents people choosing α after looking at the data.
* Nevertheless, quite a lot of people still prefer to report exact p-values.
* To many people, the advantage of allowing the reader to make up their own mind about how to interpret p = .06 outweighs any disadvantages.
* In practice, however, even among researchers who prefer exact p-values, it is quite common to just write p < .001 instead of reporting an exact value for small p, in part b/c a lot of software doesn’t actually print out the p-value when it’s that small + in part b/c a very small p value can be misleading
* Human mind sees a number like .0000000001 + it’s hard to suppress the gut feeling the evidence in favor of the alternative is a near certainty
* That’s usually wrong as life is a big, messy, complicated thing + **every statistical test ever invented relies on simplifications, approximations, + assumptions**.
* As a consequence, probably not reasonable to walk away from any statistical analysis w/ a feeling of confidence stronger than p < .001 implies.
* p < .001 is really code for “as far as *this test* is concerned, the evidence is overwhelming.”

**11.7 Running the hypothesis test in practice**

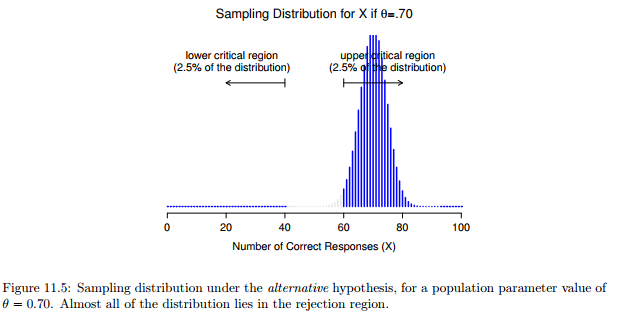
* Simplest possible problem you might ever encounter in real life; **the binomial test**, implemented by an R function called **binom.test().**
* For ESP, the p-value of 0.02 is less than the usual choice of α = .05, so you can reject the null.
* R contains a whole lot of functions corresponding to different kinds of hypothesis test.

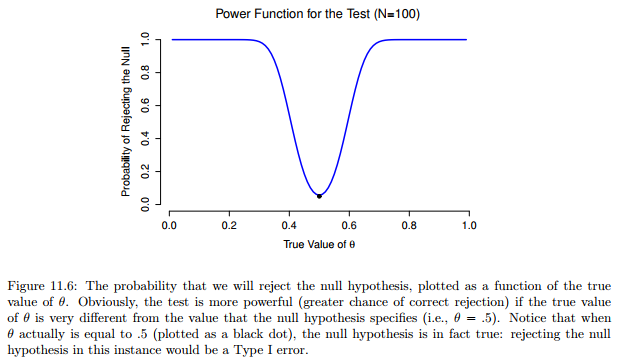
**11.8 Effect size, sample size and power**

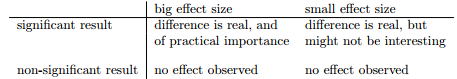
* The major design principle behind statistical hypothesis testing = we try to control Type I error rate
* Fixed α = .05 🡺 attempting to ensure only 5% of true null hypotheses are incorrectly rejected
* Doesn’t mean we don’t care about Type II errors
* Error of failing to reject the null when it is actually false is extremely annoying.
* Secondary goal of hypothesis testing = try to minimize β, the Type II error rate
* Don’t usually talk in terms of minimizing Type II errors but instead talk about maximizing **power of the test**.
* Since **power** = 1 ­- β, this is the same thing
* A Type II error occurs when the alternative is true, but we are unable to reject the null.
* Ideally, we’d be able to calculate a single number β that tells us Type II error rate, in the same way that we can set α = .05 for Type I error rate.
* Unfortunately, this is a lot trickier to do
* ESP study 🡪 alternative hypothesis corresponds to lots of possible values of θ (every value of θ except 0.5)
* Suppose the true probability of someone choosing the correct response is 55% (θ = .55).
* If so, the true sampling distribution for X is not the same one the null predicts: the most likely value for X is now 55/100.
* The whole sampling distribution has now shifted**,** while the critical regions, of course, do not change (based by definition on what the null predicts)



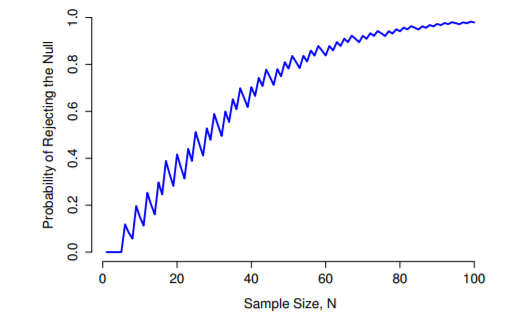
* What we’re seeing in this figure is the fact that *when the null is wrong, a much larger proportion of the sampling distribution falls in the critical region.*
* That’s what should happen: probability of rejecting the null is larger when the null is actually false
* However θ = .55 is NOT the only possibility consistent w/ the alternative
* Suppose the true value of θ is actually 0.7



* Almost the entirety of the sampling distribution has now moved into the critical region.
* Therefore, if θ = 0.7 the probability of us correctly rejecting the null (the power of the test) is much larger than if θ = 0.55.
* In short, *while θ = .55 and θ = .70 are both part of the alternative, the Type II error rate is different*
* This means is the power of a test (1 - β) depends on the true value of θ.
* 
* See the expected probability of rejecting the null for all values of θ = **power function of the test**
* It’s a summary of how good a test is, b/c it actually tells you the power (1 - β) for all possible values of θ.
* When the true value of θ is very close to 0.5, the power of the test drops very sharply, but when further away, the power is large.
* Since all models are wrong the scientist must be alert to what is *importantly* wrong
* Plot above captures a fairly basic point about hypothesis testing: If the true state of the world is very different from what the null predicts, your power will be very high; but if the true state of the world is similar to the null (but not *identical*) the power of the test is going to be very low.
* Therefore, it’s useful to be able to have some way of quantifying how “similar” the true state of the world is to the null.
* A statistic that does this is called a **measure of effect** **size**
* **Effect size** tries to capture how big the difference is between the true population parameters + the parameter values assumed by the null
* ESP 🡪 if we let θ0 = 0.5 denote the value assumed by the null, and let θ denote the true value, a simple measure of effect size could be something like the difference between the true value + null 🡺 θ - θ0
* Or possibly just the magnitude of this difference, | θ - θ0 |.
* Assume you’ve run your experiment, collected the data, + gotten a significant effect when you ran a hypothesis test.
* Isn’t it enough just to say that you’ve gotten a significant effect? Surely that’s the point of hypothesis testing? Sort of.
* The point of doing a hypothesis test is to try to demonstrate that the null is wrong, but that’s hardly the only thing we’re interested in.
* If the null claimed θ = .5, + we show it’s wrong, we’ve only really told half of the story.
* Rejecting the null implies we believe that θ != .5, but there’s a big difference between θ = .51 and θ = .8 (very wrong).
* On the other hand, if we’ve successfully rejected the null, but it looks like the true value of θ is only .51 (only be possible with a large study), the null is wrong, but it’s not at all clear we actually care, b/c the effect size is so small.
* Crude guide to understanding the relationship between statistical significance and effect sizes.
* Basically, if you don’t have a significant result, the effect size is pretty meaningless b/c you don’t have any evidence it’s even real.
* On the other hand, if you do have a significant effect but effect size is small, there’s a pretty good chance your result (although real) isn’t all that interesting.
* However, this guide is very crude + depends a lot on what exactly you’re studying.
* Small effects can be of massive practical importance in some situations



* Suppose we’re looking at differences in high school exam scores between males + females, + it turns out female scores are 1% higher on average than the males.
* If I’ve got data from thousands of students, this difference will almost certainly be statistically significant, but regardless of how small the p-value is it’s just not very interesting.
* Can’t proclaim a crisis in boys’ education on the basis of such a tiny difference
* It’s for this reason that it is becoming more standard to report some kind of standard measure of effect size along w/ results of a hypothesis test.
* Hypothesis test itself tells you whether you should believe the effect observed is real (not just due to chance) while effect size tells you whether or not you should *care*.
* Not surprisingly, scientists are fairly obsessed with maximizing the power of experiments.
* We want experiments to work, so we want to maximize the chance of rejecting the null if it is false (+ of course we usually want to believe it is false)
* 1 factor that influences power is **effect size**.
* The 1st thing you can do to increase power is to increase effect size 🡪 design study in such a way that effect size gets magnified.
* ESP 🡪 might believe psychic powers work best in quiet, darkened room w/ fewer distractions to cloud the mind.
* Therefore, try conduct experiments in just such an environment b/c if we can strengthen ESP abilities somehow, the true value of θ will go up + therefore effect size will be larger.
* In short, *clever experimental design* is 1 way to boost power b/c it can alter effect size.
* Unfortunately, often that even w/ best experimental designs, may only get a small effect.
* Perhaps ESP really does exist, but even under best conditions it’s very, very weak + under those circumstances, best bet for increasing power is to *increase sample size*.
* Although in practice, very small effect size is worrying, b/c even very minor methodological flaws might be responsible for that effect + in practice no experiment is perfect, so there are always methodological issues to worry about.
* Also notice the true population parameter θ doesn’t necessarily correspond to an *immutable fact of nature*.
* In this context, θ is just the true probability people would correctly guess a color of a card in the other room.
* As such, the population parameter can be influenced by all sorts of things (all on the assumption ESP actually exists)
* In general, more observations available = more likely you can discriminate between hypotheses.
* Run ESP experiment w/ 10 participants + 7 correctly guessed hidden card’s color, wouldn’t be terribly impressed.
* But if ran w/ 10k participants + 7k got the answer right, 🡪 much more likely to think we discovered something.
* In other words, power increases with the sample size.



* ^^ Power of a test, plotted as function of sample size n.
* In this case, true value of θ = 0.7, but the null is that θ = 0.5.
* Overall, larger n = greater power. (small zig-zags in the function occur b/c odd interactions between θ, α, + the fact that the binomial distribution is discrete (doesn’t matter for any serious purpose)
* B/c *power is important*, whenever contemplating running an experiment it’d be pretty useful to know how much power you’re likely to have.
* *It’s never possible to know for sure, since you can’t possibly know what your effect size is.*
* However, it’s often/sometimes possible to guess how big it should be.
* If so, you can guess what sample size you need = This idea is called **power analysis**
* If it’s feasible to do **power analysis**, it’s very helpful, since it can tell you something about whether you have enough time/money to be able to run the experiment successfully.
* Increasingly common to see people arguing that power analysis should be a required part of experimental design
* But, the only time scientists seems to want a power analysis in real life is when being forced to by bureaucratic process (not part of anyone’s day to day work)
* While power is an important concept, power analysis is not as useful as people make it sound, except in the rare cases where:
* (a) someone has figured out how to calculate power for your actual experimental design and
* (b) you have a pretty good idea what the effect size is likely to be.

**11.9 Some issues to consider**

* This was all the orthodox framework for **null hypothesis significance testing (NHST).**
* Understanding how NHST works is an absolute necessity, since it has been the dominant approach to inferential statistics ever since it came to prominence in the early 20th century.
* It’s what the vast majority of working scientists rely on for data analysis 🡺 need to know it.
* However, the approach is not without problems + there are a number of quirks in the framework, historical oddities in how it came to be, theoretical disputes over whether or not the framework is right, + a lot of practical traps for the unwary
* Orthodox NHST is actually a mash-up of 2 rather different approaches to hypothesis testing 🡪 by Sir Ronald Fisher + by Jerzy Neyman
* Fisher + Neyman were real people whose opinions changed over time, and at no point did either of them offer “the definitive statement” of how we should interpret their work many decades later.
* Fisher assumed you only had 1 hypothesis (null) + you find out if the null is inconsistent w/ the data
* From his perspective, check to see if data are “*sufficiently unlikely*” according to the null 🡪 that’s how Fisher defines the p-value.
* According to Fisher, if the null provided a very poor account of the data, you could safely reject it.
* But, since you don’t have any other hypotheses to compare it to, *there’s no way of “accepting the alternative*” b/c you don’t have an explicitly stated alternative.
* In contrast, Neyman thought the point of hypothesis testing was as a guide to action + his approach was somewhat more formal than Fisher’s.
* His view = there are multiple things you could do (accept the null or accept the alternative) + the point of the test was to tell you *which* one the data support.
* From this perspective, it is critical to specify alternatives properly.
* If you don’t, you don’t know how powerful the test is, or even which action makes sense.
* His framework genuinely requires a *competition between different hypotheses.*
* For Neyman, the p value didn’t directly measure probability of the data/data more extreme under the null, it was more of an abstract description about which “possible tests” were telling you to accept the null, and which “possible tests” were telling you to accept the alternative.
* What we have today is an odd mishmash of the 2.
* We talk about having both a null + an alternative (Neyman), but usually define the p-value in terms of extreme data (Fisher), but we still have α values (Neyman).
* Some statistical tests have explicitly specified alternatives (Neyman) but others are quite vague about it (Fisher).
* According to some people, we’re not allowed to talk about accepting the alternative (Fisher).
* Cannot interpret the p value as “the probability the null is true”.
* **NHST** is fundamentally a frequentist tool + as such it *does not allow you to assign probabilities to hypotheses*: **the null is either true or it is not**.
* The *Bayesian* approach to statistics interprets **probability as a degree of belief**, so it’s totally okay to say “there is a 10% chance the null is true” as a reflection of the degree of confidence you have in this hypothesis.
* You aren’t allowed to do this w/in the frequentist approach.
* If you’re a frequentist, a probability can only be defined in terms of what happens *after a large number of independent replications* (i.e., a **long run frequency**).
* If this is your interpretation of probability, talking about the “probability” that the null is true is complete gibberish: *a null hypothesis is either true or it is false*.
* There’s no way you can talk about a long run frequency for this statement.
* To talk about “the probability of the null” is as meaningless as “the color of freedom
* Most importantly, this isn’t a purely ideological matter.
* If you decide you are a Bayesian + you’re okay w/ making probability statements about hypotheses, *you have to follow the Bayesian rules for calculating those probabilities.*
* But note that *the p value is a terrible approximation to the probability that H0 is true.*
* If what you want to know is the probability of the null, the p-value is not what you’re looking for
* The theory behind hypothesis testing is a mess, + even now there are arguments in statistics about how it “should” work.
* However, disagreements among statisticians are not our real concern here. Our real concern is practical data analysis.
* While the “orthodox” approach to NHST has many drawbacks, even an unrepentant Bayesian would agree they can be useful if used responsibly.
* Most of the time they give sensible answers + you can use them to learn interesting things.
* Setting aside various ideologies + historical confusions, the fact remains that the biggest danger in all of statistics is *thoughtlessness*.
* The rush to interpret a result w/out spending time thinking through what each test actually says about the data + checking whether that’s consistent w/ how you’ve interpreted it is where the biggest trap lies
* Ex: ESP 🡪 decided to analyze data separately for male + female participants.
* Of males, 33/50 guessed the color of the card correctly 🡪 A significant effect (p = .03).
* Of females, 29/50 guessed correctly 🡪 Not a significant effect (p = .32).
* Upon observing this, it is extremely tempting for people to start wondering WHY there is a difference between males + females in terms of psychic abilities.
* However, this is wrong.
* If you think about it, we haven’t actually run a test that *explicitly* compares males to females.
* All we’ve done is *compare males to chance* (**binomial test** was significant) + compared *females to chance* (binomial test was non-significant).
* If we want to argue there is a real difference between males + females, we should probably run test of the null that there is no difference using a different hypothesis test
* When we do, turns out we have no evidence males + females are significantly different (p = .54)
* Now do you think there’s anything fundamentally different between the 2 groups? Of course not
* What’s happened here is the data from both groups (male + female) are pretty borderline: by pure chance, 1 happened to end up on the magic side of the p = .05 line, + the other didn’t, which doesn’t actually imply males + females are different.
* This mistake is so common you should always be wary of it: **the difference between significant and not-significant is NOT evidence of a real difference**
* If you want to say there’s a difference between 2 groups, you have to test for that difference
* Data analysis can be tricky to get right.
* Think about what it is you want to test, why you want to test it, and whether or not the answers that your test gives could possibly make any sense in the real world.