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

***Part IV – Statistical Theory***

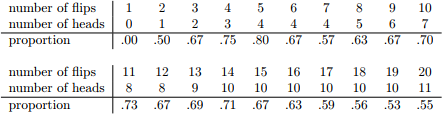
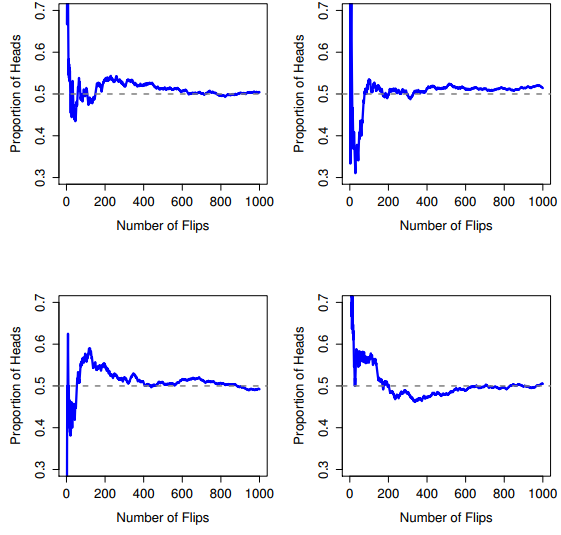
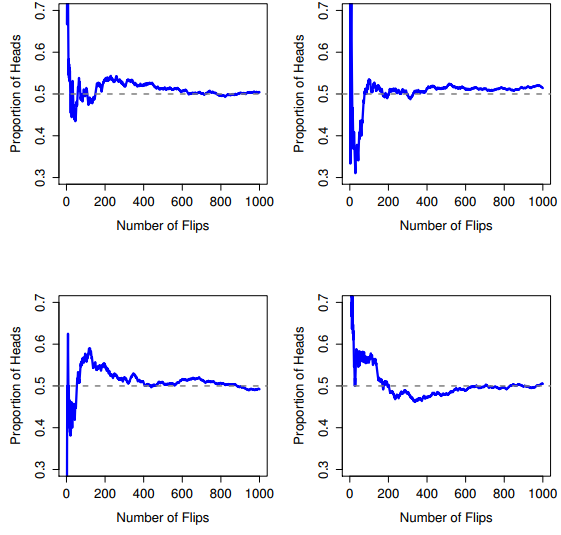
**Chapter 9 – Intro to Probability**

* Descriptive statistics is 1 of the smallest parts of statistics, + one of the least powerful.
* The bigger and more useful part of stats is that it lets you make **inferences** about data.
* Once you start thinking about stats in these terms (*stats is there to help us draw inferences from data*), you start seeing examples of it everywhere.
* For instance: *“I have a tough job,” the Premier said in response to a poll which found her government is now the most unpopular Labor administration in polling history, w/ a primary vote of just 23%.*
* This kind of remark is entirely unremarkable in papers or in everyday life, but let’s have a think about what it entails.
* A polling company has conducted a survey (usually a pretty big one b/c they can afford it)
* Imagine that they called 1K NSW voters at random, + 230 (23%) claimed they intended to vote for the ALP.
* For the 2010 Federal election, the AEC reported 4,610,795 enrolled voters in NSW; so opinions of the remaining 4,609,795 voters (99.98% of voters) remain unknown to us
* Even assuming no-one lied to the polling company, the only thing we can say **w/ 100% confidence** is that the true ALP primary vote is somewhere between 230/4610795 (~0.005%) + 4610025/4610795 (~99.83%).
* So, on what basis is it legitimate for the polling company, the newspaper, + the readership to conclude that the ALP primary vote is only about 23%?
* The answer to the question is pretty obvious: if I call 1K people at random, + 230 of them say they intend to vote for the ALP, it seems very unlikely that these are the only 230 people out of the entire voting public who actually intend to do so.
* In other words, we *assume the data collected by the polling company is pretty representative of the population at large*.
* But HOW representative? Would we be surprised to discover that the true ALP primary vote is actually 24%? 29%? 37%?
* At this point everyday intuition starts to break down a bit.
* No-one would be surprised by 24%, + everybody would be surprised by 37%, but it’s a bit hard to say whether 29% is plausible.
* We need some more powerful tools than just looking at the numbers + guessing.
* **Inferential stats** provides the tools we need to answer these sorts of questions, + since these kinds of questions lie at the heart of the scientific enterprise, they take up the lion’s share of every introductory course on stats + research methods.
* However, the theory of statistical inference is built on top of **probability theory**

**Chapter 9.1 – How are probability + stats different?**

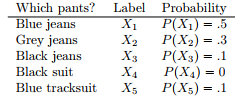
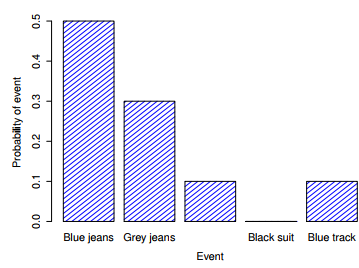
* The 2 disciplines are closely related but not identical.
* **Probability theory** = “doctrine of chances” = a branch of mathematics that tells you how often different kinds of events will happen.
* What are the chances of a fair coin coming up heads 10 times in a row?
* If I roll two six sided dice, how likely is it that I’ll roll two sixes?
* How likely is it that five cards drawn from a perfectly shuffled deck will all be hearts?
* What are the chances that I’ll win the lottery?
* Notice that all of these questions have something in common.
* In each case the “**truth of the world**” is known, + my question relates to the “what kind of events” will happen.
* In the 1st question, I know the coin is fair, so there’s a 50% chance any individual coin flip will come up heads. In the 2nd question, I know the chance of rolling a 6 on a single die is 1/6. In the 3rd question I know the deck is shuffled properly. In the 4th question, I know the lottery follows specific rules
* The critical point is that **probabilistic questions start w/ a *known model of the world***, + we use that model to do some calculations.
* The underlying model can be quite simple.
* For coin flipping, we can write down the model like this: **P(heads) = 0.5**
* When using this probability model to answer the 1st question, I don’t actually know exactly what’s going to happen.
* Maybe I’ll get 10 heads, maybe 3
* That’s the key thing: **In probability theory, the model is known, but the data are not.**
* *Statistical questions work the other way around*.
* In stats, we do NOT know the *truth about the world*, all we have is the data, + it is *from the data* that we want to learn the truth about the world.
* Statistical questions tend to look more like these:
* If my friend flips a coin 10 times + gets 10 heads, are they playing a trick on me?
* If five cards off the top of the deck are all hearts, how likely is it that the deck was shuffled?
* If the lottery commissioner’s spouse wins the lottery, how likely is it that the lottery was rigged?
* This time around, the only thing we have are data.
* I only know I saw my friend flip the coin 10 times + it came up heads every time, + what I *want to infer* is whether or not I should conclude if what I just saw was actually a fair coin being flipped 10 times in a row, or whether I should suspect that my friend is playing a trick on me.
* The data I have look like this: H H H H H H H H H H H
* What I’m trying to do is work out which “model of the world” I should put my trust in.
* If the coin is fair, the model I should adopt is one that says “the probability of heads is 0.5”
* If the coin is NOT fair, I should conclude the probability of heads is NOT 0.5 (P(heads <> 0.5 or != 0.5)
* In other words, the **statistical inference problem is to figure out *which* probability model is right**.
* Clearly, the statistical question isn’t the same as the probability question, but they’re deeply connected to one another.
* B/c of this, a good introduction to statistical theory will start w/ a discussion of what probability is + how it works.

**Chapter 9.2 - What does probability mean?**

* It might seem surprising, but while statisticians + mathematicians (mostly) agree on what the rules of probability are, there’s much less of a consensus on what the word really means.
* Suppose I want to bet on a soccer game between 2 teams of robots, + after thinking about it, I decide there is an 80% probability of Arsenal winning.
* What do I mean by that? Here are 3 possibilities...
* They’re robot teams, so I can make them play over + over again, + if I did that, Arsenal would win 8 out of every 10 games on average (**frequentist**)
* For any given game, I would only agree that betting on this game is only “fair” if a $1 bet on Milan gives a $5 payoff (i.e. I get my $1 back plus a $4 reward for being correct), as would a $4 bet on Arsenal (i.e., my $4 bet plus a $1 reward) (**Bayesian**)
* My subjective “belief” or “confidence” in an Arsenal victory is 4X as strong as my belief in a Milan victory.
* Each of these seems sensible. However they’re *not identical*, + not every statistician would endorse all of them.
* The reason is that there are different statistical ideologies + depending on which one you subscribe to, you might say that some of those statements are meaningless or irrelevant
* There are 2 main approaches that exist in the
* **The Frequentist View**
* This is more dominant one in stats + defines probability as a **long-run frequency**.
* Suppose we were to try flipping a fair coin (P(h) = 0.5), over + over again
* What might we observe? 1 possibility is that 11 of the 1st 20 coin flips (55%) came up heads.
* Now suppose that I’d been keeping a running tally of heads (which I’ll call Nh) I’ve seen, across the first N flips, + calculate the proportion of heads Nh/N every time.
* 
* Notice at the start of the sequence the proportion of heads fluctuates wildly, starting at .00 + rising as high as .80.
* Later on, one gets the impression that it dampens out a bit, w/ more + more of values actually being pretty close to the “right” answer of 0.50.
* This is the frequentist definition of probability in a nutshell: flip a fair coin over + over again, + as N grows large (approaches infinity), the proportion of heads will **converge** to 50%.
* There are some subtle technicalities mathematicians care about, but qualitatively speaking, that’s how the frequentists define probability.
* Unfortunately, I don’t have an infinite number of coins, or the infinite patience required to flip a coin an infinite number of times.
* However, I do have a CPU which excels at mindless repetitive tasks.
* 
* So I asked my computer 4 times to simulate flipping a coin 1000
* As you can see, the proportion of observed heads eventually stops fluctuating, + settles down + when it does, the number at which it finally settles is the “true” probability of heads
* The frequentist definition of probability has some desirable characteristics:
* Firstly, it is objective: the probability of an event is *necessarily grounded* in the world.
* The only way that probability statements can make sense is if they refer to (a sequence of) events that occur in the physical universe.1
* Secondly, it is unambiguous: any 2 people watching the same sequence of events unfold, trying to calculate the probability of an event, must inevitably come up w/ the same answer.
* However, it also has undesirable characteristics.
* Firstly, infinite sequences don’t exist in the physical world.
* More seriously, the frequentist definition has a narrow scope.
* There are lots of things out there human beings are happy to assign probability to in everyday language, but cannot (even in theory) be mapped onto a hypothetical sequence of events.
* For instance, if a meteorologist comes on TV + says, “the probability of rain in Adelaide on 2 November 2048 is 60%,” we humans are happy to accept this.
* But it’s not clear how to define this in frequentist terms.
* There’s only 1 city of Adelaide, + only 2 November 2048 + there’s no infinite sequence of events here, just a once-off thing.
* *Frequentist probability genuinely forbids us from making probability statements about a single event*.
* From the frequentist perspective, it will either rain tomorrow or it will not
* *There is no “probability” that attaches to a single non-repeatable event.*
* Now, there are some very clever tricks frequentists can use to get around this.
* 1 possibility is that what the meteorologist *means* is something like this: “There is a category of days for which I predict a 60% chance of rain; if we look only across those days for which I make this prediction, then on 60% of those days it will actually rain”.
* It’s weird + counterintuitive to think of it this way, but you see frequentists do it sometimes
* This doesn’t mean frequentists can’t make hypothetical statements, of course
* It’s just that if you want to make a statement about probability, then it must be possible to re-describe that statement in terms of a sequence of potentially observable events, + the relative frequencies of different outcomes that appear w/in that sequence.
* **The Bayesian view**
* Often called the subjectivist view, + it is a minority view among statisticians, but one that has been steadily gaining traction for the last several decades.
* There’re many flavors of “Bayesianism”, making hard to say exactly what a Bayesian view is
* The most common way of thinking about **subjective probability** is to define probability of an event as the ***degree of belief* that an intelligent + rational agent assigns to that truth of that event.**
* From that perspective, probabilities don’t exist in the world, *but rather in the thoughts* + *assumptions* of people + other intelligent beings.
* However, in order for this approach to work, we need some way of operationalizing “degree of belief”.
* 1 way you can do so is to formalize it in terms of “*rational gambling*”, though there are many other ways.
* Suppose I believe that there’s a 60% probability of rain tomorrow.
* If someone offers me a bet if it rains tomorrow, I win $5, + if it doesn’t I lose $5.
* Clearly, from my perspective, this is a pretty good bet.
* On the other hand, if I think the probability of rain is only 40%, it’s a bad bet to take.
* Thus, we can operationalize the notion of a *subjective* probability in terms of what bets I’m willing to accept.
* The main advantage of a Bayesian approach is:
* It allows you to assign probabilities to *any event you want to*.
* You don’t need to be limited to those events that are repeatable.
* The main disadvantage (to many people) is:
* We can’t be purely objective
* Specifying a probability requires us to specify an *entity* that has the relevant degree of belief.
* This entity might be a human, an alien, a robot, or even a statistician, but there has to be an intelligent agent out there that believes in things.
* To many people this is uncomfortable: it seems to make probability arbitrary.
* While the Bayesian approach does require that the agent in question be rational (i.e., obey the rules of probability), it does allow everyone to have their own beliefs
* I can believe the coin is fair + you don’t have to, even though we’re both rational.
* The frequentist view *doesn’t* allow any 2 observers to attribute different probabilities to the same event:
* When that happens, then at least 1 of them must be wrong.
* The Bayesian view does not prevent this from occurring.
* 2 observers w/ different background knowledge can legitimately hold different beliefs about the same event.
* In short, where the frequentist view is sometimes considered to be too narrow (forbids lots of things we want to assign probabilities to), the Bayesian view is sometimes thought to be too broad (allows too many differences between observers).
* Now that you’ve seen these 2 views independently, it’s useful to make sure you can compare the 2.
* Go back to the hypothetical soccer game. What do you think a frequentist + a Bayesian would say about these 3 statements?
* Which statement would a frequentist say is the correct definition of probability? Which one would a Bayesian? Would some of these statements be meaningless to a frequentist OR a Bayesian?
* If you’ve understood the 2 perspectives, you should have some sense of how to answer those questions.
* But which of them is right? There is probably not a right answer.
* There’s nothing mathematically incorrect about the way frequentists think about sequences of events, + there’s nothing mathematically incorrect about the way Bayesians define the beliefs of a rational agent.
* In fact, when you dig down into the details, Bayesians + frequentists actually agree about a lot of things.
* Many frequentist methods lead to decisions that Bayesians agree a rational agent would make.
* Many Bayesian methods have very good frequentist properties.
* For the most part, some are pragmatists so they’ll use any statistical method they can trust + as it turns out, that makes them prefer Bayesian methods
* Not everyone is quite so relaxed.
* Consider Sir Ronald Fisher, 1 of the towering figures of 20th century stats + a vehement opponent to all things Bayesian, whose paper on the mathematical foundations of stats referred to Bayesian probability as “an impenetrable jungle [that] arrests progress towards precision of statistical concepts”
* Or the psychologist Paul Meehl, who suggests that relying on frequentist methods could turn you into “a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring” (Meehl, 1967, p. 114).
* ***NOTE:*** The majority of statistical analyses are based on the frequentist approach.

**Chapter 9.3 - Basic Probability Theory**

* Ideological arguments between Bayesians + frequentists not w/standing, it turns out people mostly agree on the rules that probabilities should obey.
* There are lots of different ways of arriving at these rules.
* The most commonly used approach is based on the work of Andrey Kolmogorov, 1 of the great Soviet mathematicians of the 20th century. I won’t go into a lot of detail, but I’ll try to give you a
* Ex: 5 pairs of pants: 3 pairs of jeans, bottom half of a suit, + a pair of tracksuit pants 🡺 X1-X5
* Using the language of probability theory, refer to each pair of pants (each X) as an **elementary event**
* The key characteristic of elementary **events** is that every time we make an observation, the outcome will be *1 + only 1 of these events*
* Similarly, the *set of all possible events* is called a **sample space**.
* Okay, now that we have a sample space (a wardrobe) built from lots of possible events (pants)
* We want to do is assign a probability of 1 of these elementary events.
* For an event X, the probability of that event P(X) is a number that lies between 0 + 1 + the bigger the value of P(X), the more likely the event is to occur.
* So, for example, if P(X) = 0, event X is impossible + if P(X) = 1 event X is certain to occur, if P(X) = 0.5 I wear those pants half of the time.
* The last thing we need to recognize is that “*something* ALWAYS happens”.
* Every time I put on pants, I really do end up wearing pants
* What this somewhat trite statement means, in probabilistic terms, is that *the probabilities of the elementary events need to add up to 1* **=** **The Law of Total Probability**
* More importantly, if these requirements are satisfied, then we have a **probability distribution**

* If we add up the probability of all events, they sum to 1
* ***NOTE:*** Probability Theory allows you to talk about **non-elementary events** as well as elementary ones
* In the pants example, it’s perfectly legitimate to refer to the probability I wear jeans.
* In this scenario, the “I wear jeans” event is said to have happened as long as the elementary event that *actually did occur* is 1 of the *appropriate* ones; in this case “blue”, “black” or “grey jeans”.
* In mathematical terms, we defined the “jeans” event **E** to correspond to the set of elementary events (X1, X2, X3)
* If *any* of these elementary events occurs, then **E** is also said to have occurred.
* Having decided to write down the definition of the E this way, it’s pretty straightforward to state what the P(E) is: we just add everything up 🡪0.5, 0.3 + 0.1 = 0.9.
* At this point you might be thinking that this is all terribly obvious + simple + you’d be right.
* All we’ve really done is wrap some basic mathematics around a few common sense intuitions. However,
* from these simple beginnings it’s possible to construct some extremely powerful mathematical tools. I’m
* definitely not going to go into the details in this book, but what I will do is list – in Table 9.1 – some of
* the other rules that probabilities satisfy. These rules can be derived from the simple assumptions that
* I’ve outlined above, but since we don’t actually use these rules for anything in this book, I won’t do so
* here.
* 9.4
* The binomial distribution
* As you might imagine, probability distributions vary enormously, + there’s an enormous range of
* distributions out there. However, they aren’t all equally important. In fact, the vast majority of the
* content in this book relies on one of five distributions: the binomial distribution, the normal distribution,
* the t distribution, the χ
* 2
* (“chi-square”) distribution + the F distribution. Given this, what I’ll do over
* the next few sections is provide a brief introduction to all five of these, paying special attention to the
* binomial + the normal. I’ll start w/ the binomial distribution, since it’s the simplest of the five.
* 9.4.1 Introducing the binomial
* The theory of probability originated in the attempt to describe how games of chance work, so it seems
* fitting that our discussion of the binomial distribution should involve a discussion of rolling dice +
* flipping coins. Let’s imagine a simple “experiment”: in my hot little hand I’m holding 20 identical sixsided
* dice. On one face of each die there’s a picture of a skull; the other five faces are all blank. If I
* - 283 -
* proceed to roll all 20 dice, what’s the probability that I’ll get exactly 4 skulls? Assuming that the dice are
* fair, we know that the chance of any one die coming up skulls is 1 in 6; to say this another way, the skull
* probability for a single die is approximately .167. This is enough information to answer our question, so
* let’s have a look at how it’s done.
* As usual, we’ll want to introduce some names + some notation. We’ll let N denote the number of
* dice rolls in our experiment; which is often referred to as the size parameter of our binomial distribution.
* We’ll also use θ to refer to the the probability that a single die comes up skulls, a quantity that is usually
* called the success probability of the binomial.2 Finally, we’ll use X to refer to the results of our
* experiment, namely the number of skulls I get when I roll the dice. Since the actual value of X is due
* to chance, we refer to it as a random variable. In any case, now that we have all this terminology
* + notation, we can use it to state the problem a little more precisely. The quantity that we want to
* calculate is the probability that X “ 4 given that we know that θ “ .167 + N “ 20. The general
* “form” of the thing I’m interested in calculating could be written as
* PpX | θ, Nq
* + we’re interested in the special case where X “ 4, θ “ .167 + N “ 20. There’s only one more
* piece of notation I want to refer to before moving on to discuss the solution to the problem. If I want to
* say that X is generated randomly from a binomial distribution w/ parameters θ + N, the notation I
* would use is as follows:
* X „ Binomialpθ, Nq
* Yeah, yeah. I know what you’re thinking: notation, notation, notation. Really, who cares? Very
* few readers of this book are here for the notation, so I should probably move on + talk about how to
* use the binomial distribution. I’ve included the formula for the binomial distribution in Table 9.2, since
* some readers may want to play w/ it themselves, but since most people probably don’t care that much
* + b/c we don’t need the formula in this book, I won’t talk about it in any detail. Instead, I just
* want to show you what the binomial distribution looks like. To that end, Figure 9.3 plots the binomial
* probabilities for all possible values of X for our dice rolling experiment, from X “ 0 (no skulls) all the
* way up to X “ 20 (all skulls). Note that this is basically a bar chart, + is no different to the “pants
* probability” plot I drew in Figure 9.2. On the horizontal axis we have all the possible events, + on the
* vertical axis we can read off the probability of each of those events. So, the probability of rolling 4 skulls
* out of 20 times is about 0.20 (the actual answer is 0.2022036, as we’ll see in a moment). In other words,
* you’d expect that to happen about 20% of the times you repeated this experiment.
* 9.4.2 Working w/ the binomial distribution in R
* Although some people find it handy to know the formulas in Table 9.2, most people just want to know
* how to use the distributions w/out worrying too much about the maths. To that end, R has a function
* called dbinom() that calculates binomial probabilities for us. The main arguments to the function are
* • x. This is a number, or vector of numbers, specifying the outcomes whose probability you’re trying
* to calculate.
* • size. This is a number telling R the size of the experiment.
* • prob. This is the success probability for any one trial in the experiment.
* 2Note that the term “success” is pretty arbitrary, + doesn’t actually imply that the outcome is something to be desired.
* If θ referred to the probability that any one passenger gets injured in a bus crash, I’d still call it the success probability,
* but that doesn’t mean I want people to get hurt in bus crashes!
* - 284 -
* 0 5 10 15 20
* 0.00 0.05 0.10 0.15 0.20
* Number of skulls observed
* Probability
* Figure 9.3: The binomial distribution w/ size parameter of N “ 20 + an underlying success probability
* of θ “ 1{6. Each vertical bar depicts the probability of one specific outcome (i.e., one possible value of
* X). B/c this is a probability distribution, each of the probabilities must be a number between 0 +
* 1, + the heights of the bars must sum to 1 as well.
* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
* So, in order to calculate the probability of getting x = 4 skulls, from an experiment of size = 20 trials,
* in which the probability of getting a skull on any one trial is prob = 1/6 ... well, the command I would
* use is simply this:
* > dbinom( x = 4, size = 20, prob = 1/6 )
* [1] 0.2022036
* To give you a feel for how the binomial distribution changes when we alter the values of θ + N,
* let’s suppose that instead of rolling dice, I’m actually flipping coins. This time around, my experiment
* involves flipping a fair coin repeatedly, + the outcome that I’m interested in is the number of heads
* that I observe. In this scenario, the success probability is now θ “ 1{2. Suppose I were to flip the coin
* N “ 20 times. In this example, I’ve changed the success probability, but kept the size of the experiment
* the same. What does this do to our binomial distribution? Well, as Figure 9.4a shows, the main effect of
* this is to shift the whole distribution, as you’d expect. Okay, what if we flipped a coin N “ 100 times?
* Well, in that case, we get Figure 9.4b. The distribution stays roughly in the middle, but there’s a bit
* more variability in the possible outcomes.
* At this point, I should probably explain the name of the dbinom() function. Obviously, the “binom”
* part comes from the fact that we’re working w/ the binomial distribution, but the “d” prefix is probably
* a bit of a mystery. In this section I’ll give a partial explanation: specifically, I’ll explain why there is a
* prefix. As for why it’s a “d” specifically, you’ll have to wait until the next section. What’s going on here
* is that R actually provides four functions in relation to the binomial distribution. These four functions
* are dbinom(), pbinom(), rbinom() + qbinom(), + each one calculates a different quantity of interest.
* Not only that, R does the same thing for every probability distribution that it implements. No matter
* what distribution you’re talking about, there’s a d function, a p function, a q function + a r function.
* This is illustrated in Table 9.3, using the binomial distribution + the normal distribution as examples.
* - 285 -
* (a)
* 0 5 10 15 20
* 0.00 0.05 0.10 0.15
* Number of heads observed
* Probability
* (b)
* 0 20 40 60 80 100
* 0.00 0.02 0.04 0.06 0.08
* Number of heads observed
* Probability
* Figure 9.4: Two binomial distributions, involving a scenario in which I’m flipping a fair coin, so the
* underlying success probability is θ “ 1{2. In panel (a), we assume I’m flipping the coin N “ 20 times.
* In panel (b) we assume that the coin is flipped N “ 100 times.
* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
* - 286 -
* Table 9.2: Formulas for the binomial + normal distributions. We don’t really use these formulas for
* anything in this book, but they’re pretty important for more advanced work, so I thought it might be
* best to put them here in a table, where they can’t get in the way of the text. In the equation for the
* binomial, X! is the factorial function (i.e., multiply all whole numbers from 1 to X), + for the normal
* distribution “exp” refers to the exponential function, which we discussed in Chapter 7. If these equations
* don’t make a lot of sense to you, don’t worry too much about them.
* Binomial Normal
* PpX | θ, Nq “ N!
* X!pN ´ Xq!
* θ
* Xp1 ´ θq
* N´X ppX | µ, σq “ 1
* ?
* 2πσ
* exp ˆ
* ´
* pX ´ µq
* 2
* 2σ
* 2
* ˙
* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
* Table 9.3: The naming system for R probability distribution functions. Every probability distribution
* implemented in R is actually associated w/ four separate functions, + there is a pretty standardised
* way for naming these functions.
* what it does prefix normal distribution binomial distribution
* probability (density) of d dnorm() dbinom()
* cumulative probability of p pnorm() pbinom()
* generate random number from r rnorm() rbinom()
* quantile of q qnorm() qbinom()
* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
* Let’s have a look at what all four functions do. Firstly, all four versions of the function require you
* to specify the size + prob arguments: no matter what you’re trying to get R to calculate, it needs to
* know what the parameters are. However, they differ in terms of what the other argument is, + what
* the output is. So let’s look at them one at a time.
* • The d form we’ve already seen: you specify a particular outcome x, + the output is the probability
* of obtaining exactly that outcome. (the “d” is short for density, but ignore that for now).
* • The p form calculates the cumulative probability. You specify a particular quantile q, + it
* tells you the probability of obtaining an outcome smaller than or equal to q.
* • The q form calculates the quantiles of the distribution. You specify a probability value p, + gives
* you the corresponding percentile. That is, the value of the variable for which there’s a probability
* p of obtaining an outcome lower than that value.
* • The r form is a random number generator: specifically, it generates n random outcomes from
* the distribution.
* This is a little abstract, so let’s look at some concrete examples. Again, we’ve already covered dbinom()
* so let’s focus on the other three versions. We’ll start w/ pbinom(), + we’ll go back to the skull-dice
* example. Again, I’m rolling 20 dice, + each die has a 1 in 6 chance of coming up skulls. Suppose,
* however, that I want to know the probability of rolling 4 or fewer skulls. If I wanted to, I could use the
* dbinom() function to calculate the exact probability of rolling 0 skulls, 1 skull, 2 skulls, 3 skulls + 4
* skulls + then add these up, but there’s a faster way. Instead, I can calculate this using the pbinom()
* function. Here’s the command:
* > pbinom( q= 4, size = 20, prob = 1/6)
* [1] 0.7687492
* - 287 -
* In other words, there is a 76.9% chance that I will roll 4 or fewer skulls. Or, to put it another way, R is
* telling us that a value of 4 is actually the 76.9th percentile of this binomial distribution.
* Next, let’s consider the qbinom() function. Let’s say I want to calculate the 75th percentile of the
* binomial distribution. If we’re sticking w/ our skulls example, I would use the following command to
* do this:
* > qbinom( p = 0.75, size = 20, prob = 1/6 )
* [1] 4
* Hm. There’s something odd going on here. Let’s think this through. What the qbinom() function appears
* to be telling us is that the 75th percentile of the binomial distribution is 4, even though we saw from
* the pbinom() function that 4 is actually the 76.9th percentile. + it’s definitely the pbinom() function
* that is correct. I promise. The weirdness here comes from the fact that our binomial distribution doesn’t
* really have a 75th percentile. Not really. Why not? Well, there’s a 56.7% chance of rolling 3 or fewer
* skulls (you can type pbinom(3, 20, 1/6) to confirm this if you want), + a 76.9% chance of rolling 4
* or fewer skulls. So there’s a sense in which the 75th percentile should lie “in between” 3 + 4 skulls.
* But that makes no sense at all! You can’t roll 20 dice + get 3.9 of them come up skulls. This issue
* can be handled in different ways: you could report an in between value (or interpolated value, to use
* the technical name) like 3.9, you could round down (to 3) or you could round up (to 4). The qbinom()
* function rounds upwards: if you ask for a percentile that doesn’t actually exist (like the 75th in this
* example), R finds the smallest value for which the the percentile rank is at least what you asked for. In
* this case, since the “true” 75th percentile (whatever that would mean) lies somewhere between 3 + 4
* skulls, R rounds up + gives you an answer of 4. This subtlety is tedious, I admit, but thankfully it’s
* only an issue for discrete distributions like the binomial (see Section 2.2.5 for a discussion of continuous
* versus discrete). The other distributions that I’ll talk about (normal, t, χ
* 2 + F) are all continuous,
* + so R can always return an exact quantile whenever you ask for it.
* Finally, we have the random number generator. To use the rbinom() function, you specify how many
* times R should “simulate” the experiment using the n argument, + it will generate random outcomes
* from the binomial distribution. So, for instance, suppose I were to repeat my die rolling experiment 100
* times. I could get R to simulate the results of these experiments by using the following command:
* > rbinom( n = 100, size = 20, prob = 1/6 )
* [1] 3 4 8 4 4 3 1 4 4 1 3 0 3 3 4 3 1 2 2 3 4 2
* [23] 2 1 2 4 4 3 5 4 2 3 1 2 7 4 2 5 2 3 0 2 3 3
* [45] 2 2 2 3 4 4 2 0 2 4 4 3 4 1 2 3 5 3 7 5 0 5
* [67] 1 5 4 3 4 4 1 5 4 4 3 2 3 3 4 5 0 5 1 4 7 2
* [89] 5 1 1 2 4 5 5 3 3 3 3 3
* As you can see, these numbers are pretty much what you’d expect given the distribution shown in
* Figure 9.3. Most of the time I roll somewhere between 1 to 5 skulls. There are a lot of subtleties
* associated w/ random number generation using a computer,3 but for the purposes of this book we
* don’t need to worry too much about them.
* 3Since computers are deterministic machines, they can’t actually produce truly random behaviour. Instead, what they
* do is take advantage of various mathematical functions that share a lot of similarities w/ true randomness. What this
* means is that any random numbers generated on a computer are pseudorandom, + the quality of those numbers depends
* on the specific method used. By default R uses the “Mersenne twister” method. In any case, you can find out more by
* typing ?Random, but as usual the R help files are fairly dense.
* - 288 -
* −3 −2 −1 0 1 2 3
* 0.0 0.1 0.2 0.3 0.4
* Observed Value
* Probability Density
* Figure 9.5: The normal distribution w/ mean µ “ 0 + standard deviation σ “ 1. The x-axis corresponds
* to the value of some variable, + the y-axis tells us something about how likely we are to
* observe that value. However, notice that the y-axis is labelled “Probability Density” + not “Probability”.
* There is a subtle + somewhat frustrating characteristic of continuous distributions that makes
* the y axis behave a bit oddly: the height of the curve here isn’t actually the probability of observing a
* particular x value. On the other hand, it is true that the heights of the curve tells you which x values
* are more likely (the higher ones!). (see Section 9.5.1 for all the annoying details)
* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
* 9.5
* The normal distribution
* While the binomial distribution is conceptually the simplest distribution to understand, it’s not the most
* important one. That particular honour goes to the normal distribution, which is also referred to as
* “the bell curve” or a “Gaussian distribution”. A normal distribution is described using two parameters,
* the mean of the distribution µ + the standard deviation of the distribution σ. The notation that we
* sometimes use to say that a variable X is normally distributed is as follows:
* X „ Normalpµ, σq
* Of course, that’s just notation. It doesn’t tell us anything interesting about the normal distribution itself.
* As was the case w/ the binomial distribution, I have included the formula for the normal distribution
* in this book, b/c I think it’s important enough that everyone who learns stats should at least
* look at it, but since this is an introductory text I don’t want to focus on it, so I’ve tucked it away
* in Table 9.2. Similarly, the R functions for the normal distribution are dnorm(), pnorm(), qnorm() +
* rnorm(). However, they behave in pretty much exactly the same way as the corresponding functions for
* the binomial distribution, so there’s not a lot that you need to know. The only thing that I should point
* out is that the argument names for the parameters are mean + sd. In pretty much every other respect,
* there’s nothing else to add.
* - 289 -
* 0 2 4 6 8 10
* 0.0 0.1 0.2 0.3 0.4
* Observed Value
* Probability Density
* Figure 9.6: An illustration of what happens when you change the mean of a normal distribution. The solid
* line depicts a normal distribution w/ a mean of µ “ 4. The dashed line shows a normal distribution w/
* a mean of µ “ 7. In both cases, the standard deviation is σ “ 1. Not surprisingly, the two distributions
* have the same shape, but the dashed line is shifted to the right.
* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
* Instead of focusing on the maths, let’s try to get a sense for what it means for a variable to be
* normally distributed. To that end, have a look at Figure 9.5, which plots a normal distribution w/
* mean µ “ 0 + standard deviation σ “ 1. You can see where the name “bell curve” comes from: it
* looks a bit like a bell. Notice that, unlike the plots that I drew to illustrate the binomial distribution, the
* picture of the normal distribution in Figure 9.5 shows a smooth curve instead of “histogram-like” bars.
* This isn’t an arbitrary choice: the normal distribution is continuous, whereas the binomial is discrete.
* For instance, in the die rolling example from the last section, it was possible to get 3 skulls or 4 skulls,
* but impossible to get 3.9 skulls. The figures that I drew in the previous section reflected this fact: in
* Figure 9.3, for instance, there’s a bar located at X “ 3 + another one at X “ 4, but there’s nothing
* in between. Continuous quantities don’t have this constraint. For instance, suppose we’re talking about
* the weather. The temperature on a pleasant Spring day could be 23 degrees, 24 degrees, 23.9 degrees, or
* anything in between since temperature is a continuous variable, + so a normal distribution might be
* quite appropriate for describing Spring temperatures.4
* W/ this in mind, let’s see if we can’t get an intuition for how the normal distribution works. Firstly,
* let’s have a look at what happens when we play around w/ the parameters of the distribution. To
* that end, Figure 9.6 plots normal distributions that have different means, but have the same standard
* deviation. As you might expect, all of these distributions have the same “width”. The only difference
* between them is that they’ve been shifted to the left or to the right. In every other respect they’re
* identical. In contrast, if we increase the standard deviation while keeping the mean constant, the peak
* 4
* In practice, the normal distribution is so handy that people tend to use it even when the variable isn’t actually
* continuous. As long as there are enough categories (e.g., Likert scale responses to a questionnaire), it’s pretty standard
* practice to use the normal distribution as an approximation. This works out much better in practice than you’d think.
* - 290 -
* 0 2 4 6 8 10
* 0.0 0.1 0.2 0.3 0.4
* Observed Value
* Probability Density
* Figure 9.7: An illustration of what happens when you change the the standard deviation of a normal
* distribution. Both distributions plotted in this figure have a mean of µ “ 5, but they have different
* standard deviations. The solid line plots a distribution w/ standard deviation σ “ 1, + the dashed
* line shows a distribution w/ standard deviation σ “ 2. As a consequence, both distributions are
* “centred” on the same spot, but the dashed line is wider than the solid one.
* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
* of the distribution stays in the same place, but the distribution gets wider, as you can see in Figure 9.7.
* Notice, though, that when we widen the distribution, the height of the peak shrinks. This has to happen:
* in the same way that the heights of the bars that we used to draw a discrete binomial distribution have
* to sum to 1, the total area under the curve for the normal distribution must equal 1. Before moving
* on, I want to point out one important characteristic of the normal distribution. Irrespective of what
* the actual mean + standard deviation are, 68.3% of the area falls w/in 1 standard deviation of the
* mean. Similarly, 95.4% of the distribution falls w/in 2 standard deviations of the mean, + 99.7% of
* the distribution is w/in 3 standard deviations. This idea is illustrated in Figure 9.8.
* 9.5.1 Probability density
* There’s something I’ve been trying to hide throughout my discussion of the normal distribution,
* something that some introductory textbooks omit completely. They might be right to do so: this “thing”
* that I’m hiding is weird + counterintuitive even by the admittedly distorted standards that apply in
* stats. Fortunately, it’s not something that you need to understand at a deep level in order to do
* basic stats: rather, it’s something that starts to become important later on when you move beyond
* the basics. So, if it doesn’t make complete sense, don’t worry: try to make sure that you follow the gist
* of it.
* Throughout my discussion of the normal distribution, there’s been one or two things that don’t quite
* make sense. Perhaps you noticed that the y-axis in these figures is labelled “Probability Density” rather
* - 291 -
* −4 −3 −2 −1 0 1 2 3 4
* Shaded Area = 68.3%
* −4 −3 −2 −1 0 1 2 3 4
* Shaded Area = 95.4%
* (a) (b)
* Figure 9.8: The area under the curve tells you the probability that an observation falls w/in a particular
* range. The solid lines plot normal distributions w/ mean µ “ 0 + standard deviation σ “ 1. The
* shaded areas illustrate “areas under the curve” for two important cases. In panel a, we can see that there
* is a 68.3% chance that an observation will fall w/in one standard deviation of the mean. In panel b, we
* see that there is a 95.4% chance that an observation will fall w/in two standard deviations of the mean.
* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
* −4 −3 −2 −1 0 1 2 3 4
* Shaded Area = 15.9%
* −4 −3 −2 −1 0 1 2 3 4
* Shaded Area = 34.1%
* (a) (b)
* Figure 9.9: Two more examples of the “area under the curve idea”. There is a 15.9% chance that an
* observation is one standard deviation below the mean or smaller (panel a), + a 34.1% chance that
* the observation is greater than one standard deviation below the mean but still below the mean (panel
* b). Notice that if you add these two numbers together you get 15.9% ` 34.1% “ 50%. For normally
* distributed data, there is a 50% chance that an observation falls below the mean. + of course that
* also implies that there is a 50% chance that it falls above the mean.
* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
* - 292 -
* than density. Maybe you noticed that I used P(X) instead of P(X) when giving the formula for the
* normal distribution. Maybe you’re wondering why R uses the “d” prefix for functions like dnorm(). +
* maybe, just maybe, you’ve been playing around w/ the dnorm() function, + you accidentally typed
* in a command like this:
* > dnorm( x = 1, mean = 1, sd = 0.1 )
* [1] 3.989423
* + if you’ve done the last part, you’re probably very confused. I’ve asked R to calculate the probability
* that x = 1, for a normally distributed variable w/ mean = 1 + standard deviation sd = 0.1; + it
* tells me that the probability is 3.99. But, as we discussed earlier, probabilities can’t be larger than 1. So
* either I’ve made a mistake, or that’s not a probability.
* As it turns out, the second answer is correct. What we’ve calculated here isn’t actually a probability:
* it’s something else. To understand what that something is, you have to spend a little time thinking
* about what it really means to say that X is a continuous variable. Let’s say we’re talking about the
* temperature outside. The thermometer tells me it’s 23 degrees, but I know that’s not really true. It’s
* not exactly 23 degrees. Maybe it’s 23.1 degrees, I think to myself. But I know that that’s not really true
* either, b/c it might actually be 23.09 degrees. But, I know that... well, you get the idea. The tricky
* thing w/ genuinely continuous quantities is that you never really know exactly what they are.
* Now think about what this implies when we talk about probabilities. Suppose that tomorrow’s
* maximum temperature is sampled from a normal distribution w/ mean 23 + standard deviation
* 1. What’s the probability that the temperature will be exactly 23 degrees? The answer is “zero”, or
* possibly, “a number so close to zero that it might as well be zero”. Why is this? It’s like trying to throw
* a dart at an infinitely small dart board: no matter how good your aim, you’ll never hit it. In real life
* you’ll never get a value of exactly 23. It’ll always be something like 23.1 or 22.99998 or something. In
* other words, it’s completely meaningless to talk about the probability that the temperature is exactly 23
* degrees. However, in everyday language, if I told you that it was 23 degrees outside + it turned out
* to be 22.9998 degrees, you probably wouldn’t call me a liar. B/c in everyday language, “23 degrees”
* usually means something like “somewhere between 22.5 + 23.5 degrees”. + while it doesn’t feel very
* meaningful to ask about the probability that the temperature is exactly 23 degrees, it does seem sensible
* to ask about the probability that the temperature lies between 22.5 + 23.5, or between 20 + 30, or
* any other range of temperatures.
* The point of this discussion is to make clear that, when we’re talking about continuous distributions,
* it’s not meaningful to talk about the probability of a specific value. However, what we can talk about
* is the probability that the value lies w/in a particular range of values. To find out the probability
* associated w/ a particular range, what you need to do is calculate the “area under the curve”. We’ve
* seen this concept already: in Figure 9.8, the shaded areas shown depict genuine probabilities (e.g., in
* Figure 9.8a it shows the probability of observing a value that falls w/in 1 standard deviation of the
* mean).
* Okay, so that explains part of the story. I’ve explained a little bit about how continuous probability
* distributions should be interpreted (i.e., area under the curve is the key thing), but I haven’t actually
* explained what the dnorm() function actually calculates. Equivalently, what does the formula for P(X)
* that I described earlier actually mean? Obviously, P(X) doesn’t describe a probability, but what is
* it? The name for this quantity P(X) is a probability density, + in terms of the plots we’ve been
* drawing, it corresponds to the height of the curve. The densities themselves aren’t meaningful in +
* of themselves: but they’re “rigged” to ensure that the area under the curve is always interpretable as
* - 293 -
* −4 −2 0 2 4
* 0.0 0.1 0.2 0.3 0.4
* Observed Value
* Probability Density
* Figure 9.10: A t distribution w/ 3 degrees of freedom (solid line). It looks similar to a normal distribution,
* but it’s not quite the same. For comparison purposes, I’ve plotted a standard normal distribution
* as the dashed line. Note that the “tails” of the t distribution are “heavier” (i.e., extend further outwards)
* than the tails of the normal distribution? That’s the important difference between the two.
* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
* genuine probabilities. To be honest, that’s about as much as you really need to know for now.5
* 9.6
* Other useful distributions
* The normal distribution is the distribution that stats makes most use of (for reasons to be discussed
* shortly), + the binomial distribution is a very useful one for lots of purposes. But the world of stats
* is filled w/ probability distributions, some of which we’ll run into in passing. In particular, the three
* that will appear in this book are the t distribution, the χ
* 2 distribution + the F distribution. I won’t
* give formulas for any of these, or talk about them in too much detail, but I will show you some pictures.
* • The t distribution is a continuous distribution that looks very similar to a normal distribution,
* but has heavier tails: see Figure 9.10. This distribution tends to arise in situations where you
* think that the data actually follow a normal distribution, but you don’t know the mean or standard
* deviation. As you might expect, the relevant R functions are dt(), pt(), qt() + rt(), + we’ll
* run into this distribution again in Chapter 13.
* 5For those readers who know a little calculus, I’ll give a slightly more precise explanation. In the same way that
* probabilities are non-negative numbers that must sum to 1, probability densities are non-negative numbers that must
* integrate to 1 (where the integral is taken across all possible values of X). To calculate the probability that X falls between
* a + b we calculate the definite integral of the density function over the corresponding range, ş
* b
* a
* P(X) dx. If you don’t
* remember or never learned calculus, don’t worry about this. It’s not needed for this book.
* - 294 -
* 0 2 4 6 8 10
* 0.00 0.05 0.10 0.15 0.20 0.25
* Observed Value
* Probability Density
* Figure 9.11: A χ
* 2 distribution w/ 3 degrees of freedom. Notice that the observed values must always be
* greater than zero, + that the distribution is pretty skewed. These are the key features of a chi-square
* distribution.
* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
* 0 2 4 6 8 10
* 0.0 0.1 0.2 0.3 0.4 0.5 0.6 0.7
* Observed Value
* Probability Density
* Figure 9.12: An F distribution w/ 3 + 5 degrees of freedom. Qualitatively speaking, it looks pretty
* similar to a chi-square distribution, but they’re not quite the same in general.
* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
* - 295 -
* • The χ
* 2 distribution is another distribution that turns up in lots of different places. The situation
* in which we’ll see it is when doing categorical data analysis (Chapter 12), but it’s one of those
* things that actually pops up all over the place. When you dig into the maths (+ who doesn’t love
* doing that?), it turns out that the main reason why the χ
* 2 distribution turns up all over the place
* is that, if you have a bunch of variables that are normally distributed, square their values + then
* add them up (a procedure referred to as taking a “sum of squares”), this sum has a χ
* 2 distribution.
* You’d be amazed how often this fact turns out to be useful. Anyway, here’s what a χ
* 2 distribution
* looks like: Figure 9.11. Once again, the R commands for this one are pretty predictable: dchisq(),
* pchisq(), qchisq(), rchisq().
* • The F distribution looks a bit like a χ
* 2 distribution, + it arises whenever you need to compare
* two χ
* 2 distributions to one another. Admittedly, this doesn’t exactly sound like something that
* any sane person would want to do, but it turns out to be very important in real world data analysis.
* Remember when I said that χ
* 2
* turns out to be the key distribution when we’re taking a “sum of
* squares”? Well, what that means is if you want to compare two different “sums of squares”, you’re
* probably talking about something that has an F distribution. Of course, as yet I still haven’t given
* you an example of anything that involves a sum of squares, but I will... in Chapter 14. + that’s
* where we’ll run into the F distribution. Oh, + here’s a picture: Figure 9.12. + of course we
* can get R to do things w/ F distributions just by using the commands df(), pf(), qf() + rf().
* B/c these distributions are all tightly related to the normal distribution + to each other, +
* b/c they are will turn out to be the important distributions when doing inferential stats later in
* this book, I think it’s useful to do a little demonstration using R, just to “convince ourselves” that these
* distributions really are related to each other in the way that they’re supposed to be. First, we’ll use the
* rnorm() function to generate 1000 normally-distributed observations:
* > normal.a <- rnorm( n=1000, mean=0, sd=1 )
* > print(normal.a)
* [1] 0.2913131706 -0.4156161554 0.1482611948 0.8516858463 -0.6658081840
* [6] 0.8827940964 1.3757851963 0.2497812249 -0.1926513775 0.2160192605
* [11] -0.7982884040 -1.4027212056 0.0281455244 -0.1266362460 0.8645205990
* BLAH BLAH BLAH
* So the normal.a variable contains 1000 numbers that are normally distributed, + have mean 0 +
* standard deviation 1, + the actual print out of these numbers goes on for rather a long time. Note
* that, b/c the default parameters of the rnorm() function are mean=0 + sd=1, I could have shortened
* the command to rnorm( n=1000 ). In any case, what we can do is use the hist() function to draw a
* histogram of the data, like so:
* > hist( normal.a )
* If you do this, you should see something similar to Figure 9.13a. Your plot won’t look quite as pretty as
* the one in the figure, of course, b/c I’ve played around w/ all the formatting (see Chapter 6), +
* I’ve also plotted the true distribution of the data as a solid black line (i.e., a normal distribution w/
* mean 0 + standard deviation 1) so that you can compare the data that we just generated to the true
* distribution.
* In the previous example all I did was generate lots of normally distributed observations using rnorm()
* + then compared those to the true probability distribution in the figure (using dnorm() to generate the
* black line in the figure, but I didn’t show the commmands for that). Now let’s try something trickier.
* We’ll try to generate some observations that follow a chi-square distribution w/ 3 degrees of freedom,
* but instead of using rchisq(), we’ll start w/ variables that are normally distributed, + see if we
* - 296 -
* Simulated Normal Data
* −4 −2 0 2 4
* Simulated Chi−Square Data
* 0 5 10 15
* (a) (b)
* Simulated t Data
* −4 −2 0 2 4
* Simulated F Data
* 0 1 2 3 4 5 6
* (c) (d)
* Figure 9.13: Data sampled from different distributions. See the main text for details.
* . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . .
* can exploit the known relationships between normal + chi-square distributions to do the work. As I
* mentioned earlier, a chi-square distribution w/ k degrees of freedom is what you get when you take
* k normally-distributed variables (w/ mean 0 + standard deviation 1), square them, + add them
* up. Since we want a chi-square distribution w/ 3 degrees of freedom, we’ll need to supplement our
* normal.a data w/ two more sets of normally-distributed observations, imaginatively named normal.b
* + normal.c:
* > normal.b <- rnorm( n=1000 ) # another set of normally distributed data
* > normal.c <- rnorm( n=1000 ) # + another!
* Now that we’ve done that, the theory says we should square these + add them together, like this
* > chi.sq.3 <- (normal.a)^2 + (normal.b)^2 + (normal.c)^2
* + the resulting chi.sq.3 variable should contain 1000 observations that follow a chi-square distribution
* w/ 3 degrees of freedom. You can use the hist() function to have a look at these observations yourself,
* using a command like this,
* > hist( chi.sq.3 )
* + you should obtain a result that looks pretty similar to the plot in Figure 9.13b. Once again, the plot
* that I’ve drawn is a little fancier: in addition to the histogram of chi.sq.3, I’ve also plotted a chi-square
* distribution w/ 3 degrees of freedom. It’s pretty clear that – even though I used rnorm() to do all the
* - 297 -
* work rather than rchisq() – the observations stored in the chi.sq.3 variable really do follow a chi-square
* distribution. Admittedly, this probably doesn’t seem all that interesting right now, but later on when we
* start encountering the chi-square distribution in Chapter 12, it will be useful to understand the fact that
* these distributions are related to one another.
* We can extend this demonstration to the t distribution + the F distribution. Earlier, I implied that
* the t distribution is related to the normal distribution when the standard deviation is unknown. That’s
* certainly true, + that’s the what we’ll see later on in Chapter 13, but there’s a somewhat more precise
* relationship between the normal, chi-square + t distributions. Suppose we “scale” our chi-square data
* by dividing it by the degrees of freedom, like so
* > scaled.chi.sq.3 <- chi.sq.3 / 3
* We then take a set of normally distributed variables + divide them by (the square root of) our scaled
* chi-square variable which had df “ 3, + the result is a t distribution w/ 3 degrees of freedom:
* > normal.d <- rnorm( n=1000 ) # yet another set of normally distributed data
* > t.3 <- normal.d / sqrt( scaled.chi.sq.3 ) # divide by square root of scaled chi-square to get t
* If we plot the histogram of t.3, we end up w/ something that looks very similar to Figure 9.13c.
* Similarly, we can obtain an F distribution by taking the ratio between two scaled chi-square distributions.
* Suppose, for instance, we wanted to generate data from an F distribution w/ 3 + 20 degrees of freedom.
* We could do this using df(), but we could also do the same thing by generating two chi-square variables,
* one w/ 3 degrees of freedom, + the other w/ 20 degrees of freedom. As the example w/ chi.sq.3
* illustrates, we can actually do this using rnorm() if we really want to, but this time I’ll take a short cut:
* > chi.sq.20 <- rchisq( 1000, 20) # generate chi square data w/ df = 20...
* > scaled.chi.sq.20 <- chi.sq.20 / 20 # scale the chi square variable...
* > F.3.20 <- scaled.chi.sq.3 / scaled.chi.sq.20 # take the ratio of the two chi squares...
* > hist( F.3.20 ) # ... + draw a picture
* The resulting F.3.20 variable does in fact store variables that follow an F distribution w/ 3 + 20
* degrees of freedom. This is illustrated in Figure 9.13d, which plots the histgram of the observations
* stored in F.3.20 against the true F distribution w/ df1 “ 3 + df2 “ 20. Again, they match.
* Okay, time to wrap this section up. We’ve seen three new distributions: χ
* 2
* , t + F. They’re all
* continuous distributions, + they’re all closely related to the normal distribution. I’ve talked a little
* bit about the precise nature of this relationship, + shown you some R commands that illustrate this
* relationship. The key thing for our purposes, however, is not that you have a deep understanding of all
* these different distributions, nor that you remember the precise relationships between them. The main
* thing is that you grasp the basic idea that these distributions are all deeply related to one another, + to
* the normal distribution. Later on in this book, we’re going to run into data that are normally distributed,
* or at least assumed to be normally distributed. What I want you to understand right now is that, if you
* make the assumption that your data are normally distributed, you shouldn’t be surprised to see χ
* 2
* , t
* + F distributions popping up all over the place when you start trying to do your data analysis.
* 9.7
* Summary
* In this chapter we’ve talked about probability. We’ve talked what probability means, + why statisticians
* can’t agree on what it means. We talked about the rules that probabilities have to obey. + we
* - 298 -
* introduced the idea of a probability distribution, + spent a good chunk of the chapter talking about
* some of the more important probability distributions that statisticians work w/. The section by section
* breakdown looks like this:
* • Probability theory versus stats (Section 9.1)
* • Frequentist versus Bayesian views of probability (Section 9.2)
* • Basics of probability theory (Section 9.3)
* • Binomial distribution (Section 9.4), normal distribution (Section 9.5), + others (Section 9.6)
* As you’d expect, my coverage is by no means exhaustive. Probability theory is a large branch of
* mathematics in its own right, entirely separate from its application to stats + data analysis. As
* such, there are thousands of books written on the subject + universities generally offer multiple classes
* devoted entirely to probability theory. Even the “simpler” task of documenting standard probability
* distributions is a big topic. I’ve described five standard probability distributions in this chapter, but
* sitting on my bookshelf I have a 45-chapter book called “Statistical Distributions” (M. Evans, Hastings,
* & Peacock, 2011) that lists a lot more than that. Fortunately for you, very little of this is necessary.
* You’re unlikely to need to know dozens of statistical distributions when you go out + do real world
* data analysis, + you definitely won’t need them for this book, but it never hurts to know that there’s
* other possibilities out there.
* Picking up on that last point, there’s a sense in which this whole chapter is something of a digression.
* Many undergraduate psychology classes on stats skim over this content very quickly (I know mine
* did), + even the more advanced classes will often “forget” to revisit the basic foundations of the field.
* Most academic psychologists would not know the difference between probability + density, + until
* recently very few would have been aware of the difference between Bayesian + frequentist probability.
* However, I think it’s important to understand these things before moving onto the applications. For
* example, there are a lot of rules about what you’re “allowed” to say when doing statistical inference,
* + many of these can seem arbitrary + weird. However, they start to make sense if you understand
* that there is this Bayesian/frequentist distinction. Similarly, in Chapter 13 we’re going to talk about
* something called the t-test, + if you really want to have a grasp of the mechanics of the t-test it really
* helps to have a sense of what a t-distribution actually looks like. You get the idea, I hope