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

***Part IV – Statistical Theory***

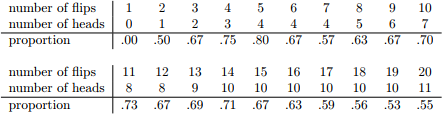
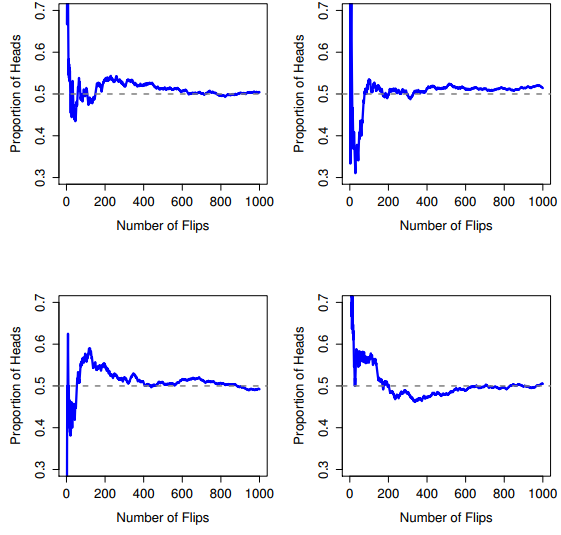
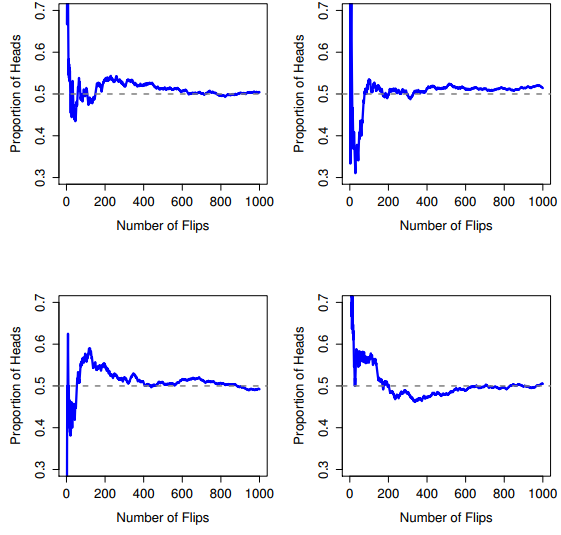
**Chapter 9 – Intro to Probability**

* 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?**

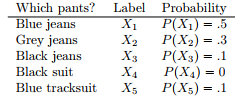
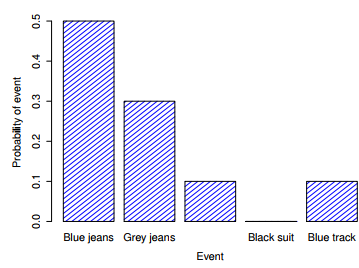
* The 2 disciplines are closely related but not identical.
* **Probability theory** = “doctrine of chances” = a branch of mathematics that tells you how often different kinds of events will happen.
* What are the chances of a fair coin coming up heads 10 times in a row?
* If I roll two six sided dice, 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?**

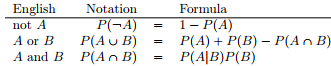
* It might seem surprising, but while statisticians + mathematicians (mostly) agree on what the rules of probability are, there’s much less of a consensus on what the word really means.
* Suppose I want to bet on a soccer game between 2 teams of robots, + after thinking about it, I decide there is an 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”
* ***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**

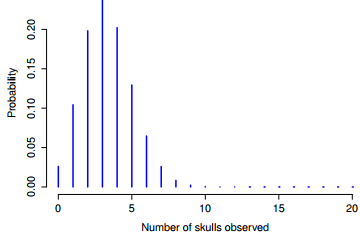
* If we add up the probability of all events, they sum to 1
* ***NOTE:*** Probability Theory allows you to talk about **non-elementary events** as well as elementary ones
* In the pants example, it’s perfectly legitimate to refer to the probability I wear jeans.
* In this scenario, the “I wear jeans” 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.



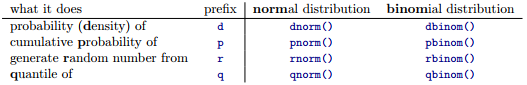
**9.4 - The Binomial Distribution**

* Probability distributions vary enormously, + there’s an enormous range of distributions out there, but they aren’t all equally important
* The theory of probability originated in the attempt to describe how games of chance work
* Ex: 20 identical 6-sided dice + on 1 face of each die there’s a picture of a skull; the other 5 are blank.
* If I proceed to roll all 20 dice, what’s the probability that I’ll get exactly 4 skulls?
* Assuming dice are fair, chance of 1 one die coming up skulls = 1 in 6 = 0.167.
* N = number of dice rolls in our experiment = the **size parameter** of our **binomial distribution**.
* θ = the probability a single die comes up skulls, a quantity = the **success probability of the binomial**
* X = the results of our experiment = 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**.
* We want to calculate the probability that X = 4 given θ = 0.167 + N = 20.
* The general form of the thing I’m interested in calculating could be written as **P(X | θ, N)**
* If I say X is generated *randomly* from a binomial distribution w/ parameters θ + N, we can write it as **X ~ Binomial(θ, N)**





* The probability of rolling 4 skulls out of 20 times is about 0.20 (0.2022036)
* In other words, you’d expect that to happen about 20% of the times you repeated this experiment.



* The **density (d) form =** specify a particular outcome x, + the output = probability of obtaining exactly that outcome
* The **(cumulative probability) p** = specify a particular quantile q + it gives the probability of obtaining an outcome <= q.
* The **(quantiles) q** form = specify a probability value p + gives the corresponding percentile (value of the variable for which there’s a probability p of obtaining an outcome lower than that value).
* The **random number generator (r)** form = generates n random outcomes from the distribution.



* In other words, there is a 76.9% chance we will roll 4 or fewer skulls.
* *A value of 4 is actually the 76.9th percentile of this binomial distribution.*



* It appears to be telling us is the 75th percentile of the binomial distribution is 4, even though we saw from pbinom()that 4 is actually the *76.9th* percentile
* *The pbinom() function is correct.*
* Binomial distribution doesn’t really have a 75th percentile. Not really

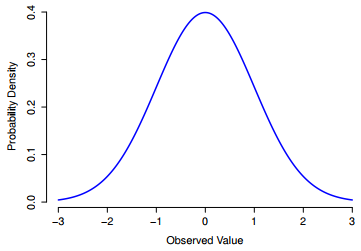


* So there’s a 56.7% chance of rolling 3 or fewer skulls + a 76.9% chance of rolling 4 or fewer skulls.
* So there’s a sense in which the 75th percentile should lie “in between” 3 + 4 skulls.
* Impossible. Can’t roll 20 dice + get 3.9 of them come up skulls.
* This issue can be handled in different ways:
* Could report an in between value (**interpolated value**) like 3.9
* Could round down to 3 or up to 4
* qbinom rounds upwards 🡪 if you ask for a percentile that doesn’t actually exist (like the 75th), R finds the smallest value for which the percentile rank *is at least* what you asked for.
* In this case, since the “true” 75th percentile lies somewhere between 3 + 4, R rounds up + gives you an answer of 4
* This only an issue for **discrete distributions** like the binomial
* Other distributions like normal, t, χ2 + F are all
* **\*\*\***Since CPUs are deterministic machines, they can’t actually produce *truly* random behavior.
* Instead, they take advantage of various mathematical functions that share a lot of similarities w/ true randomness
* Any random numbers generated on a CPU are **pseudorandom**, + the quality of those numbers depends on the specific method used
* R uses the **Mersenne twister method** by default
* Figure 9.5: The normal distribution w/ mean µ “ 0 + SD 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)

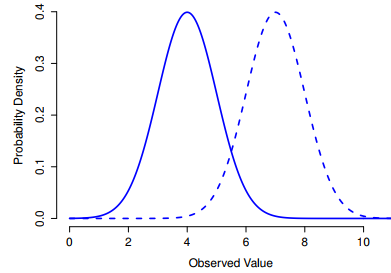
**9.5 - The Normal Distribution**

* While a binomial distribution is conceptually the simplest distribution to understand, it’s not the most important 🡪 **the normal distribution** or the **bell curve** or a **Gaussian distribution**
* We say a variable X is normally as 

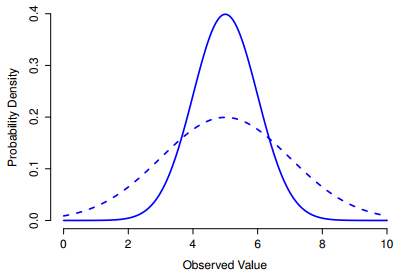




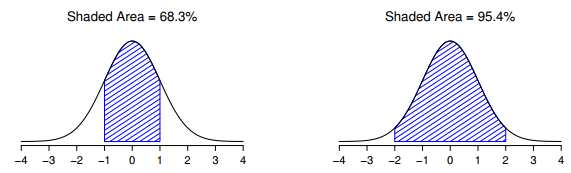
* The normal distribution = mean µ = 0 + SD σ = 1, x-axis corresponds to the value of some variable, + the y-axis tells us something about how likely we are to observe that value.
* However, notice that the y-axis is labelled **Probability Density +** *NOT*Probability
* Height of the curve here *isn’t* actually the probability of observing a particular x
* The heights *DO* tell you which x values are more likely
* The normal distribution is **continuous**, whereas the binomial is discrete (possible to get 3 skulls or 4 skulls, but impossible to get 3.9 skulls)
* Temperature could be 23 degrees, 24 degrees, 23.9 degrees, or anything in between since temperature is a continuous variable
* In practice, the normal distribution is so handy people tend to use it even when a variable isn’t actually continuous.
* As long as there are enough categories, it’s pretty standard practice to use the normal distribution as an approximation.
* This works out much better in practice than you’d think.

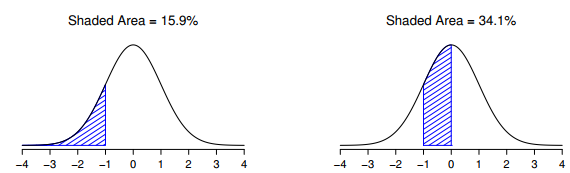


* ^^^Change the mean of a normal distribution.
* In both cases, the σ = 1 + not surprisingly, the 2 distributions have the same shape
* If we increase σ while keeping the mean constant, the peak stays in the same place, but the distribution gets wider
* Notice, though, that when we widen the distribution, the height of the peak shrinks.

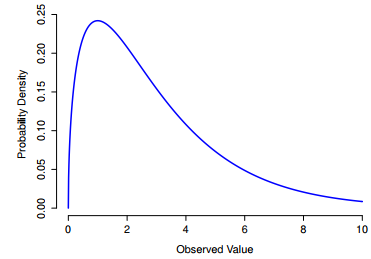


* This happens in the same way the heights of the bars used to draw a discrete binomial distribution have to sum to 1, the **total AUC for the normal distribution must equal 1**.
* Irrespective of what the actual µ + SD are, 68.3% of the area falls w/in 1 SD of µ, 95.4% of the distribution falls w/in 2 SDs of µ, + 99.7% of the distribution is w/in 3 SDs.
* Y-axis is labelled **Probability Density** rather than just *density +* we use **p(X)** instead of P(X) when giving the formula for the normal distribution
* W/ genuinely continuous quantities, you never really know exactly what they are.
* Now think about what this implies when we talk about probabilities.
* Suppose tomorrow’s max temp is sampled from a normal distribution w/ mean = 23 + SD = 1.
* What’s the probability that the temperature will be exactly 23 degrees?
* The answer is a number so close to 0 that it might as well be 0
* It’s completely meaningless to talk about the probability the temperature is exactly 23 degrees.
* In everyday language, 23 degrees usually means “somewhere between 22.5 + 23.5 degrees”
* When talking about continuous distributions, it’s not meaningful to talk about the probability of a *specific* value but we can talk about the probability a value lies w/in a particular range of values.
* *To find out the probability associated w/ a particular range, you need calculate the* ***AUC***

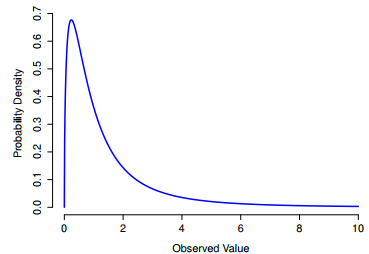




* The shaded AUCs depict genuine probabilities that an observation falls w/in a particular range.
* There is a 68.3% chance an observation falls w/in 1 SD of µ + a 95.4% chance it falls w/in 2 SDs.
* There is a 15.9% chance an observation is 1 SD below µ or smaller + a 34.1% chance an observation is greater than 1 SD below µ but still below µ
* Notice if you add these 2 together you get 50% 🡪 For normally distributed data, there is a 50% chance an observation falls below µ.
* This also implies there is a 50% chance that it falls above the mean
* So p(X) doesn’t describe probability, but **probability density** + it corresponds to the height of the curve
* The densities themselves aren’t meaningful in + of themselves, but they’re rigged to ensure the AUC is always interpretable as genuine probabilities.
* 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 (integral is taken across all possible values of X).
* To calculate the probability X falls between a-b, we calculate the definite integral of the density function over the corresponding range 
* The normal distribution is the distribution that stats makes most use of, + the binomial distribution is a very useful one for lots of purposes
* **The t distribution** = a continuous distribution that looks very similar to a normal distribution, but w/ heavier tails (extend further outward)
* tends to arise in situations where you think the data actually follow a normal distribution, but you don’t know the mean or standard deviation.
* The **χ2 distribution** turns up all over the place when doing categorical data analysis
* If you have a bunch of variables that’re normally distributed + we square their values + then add them up (**sum of squares**), this sum has a χ2 distribution
* **χ2 distribution** with k dF = k normally-distributed populations/variables, square them, + add them up.
* If we want a χ2 distribution w/ dF = 3, we’ll need 3 sets of normally-distributed observations



* **The F distribution** looks a bit like a χ2 distribution + arises whenever you need to compare 2 χ2 distributions to one another
* χ2 is the key distribution when we’re taking a sum of squares
* If you want to compare 2 different sums of squares, you’re probably talking about something that has an F distribution.



* So, the t distribution is related to the normal distribution when the *SD is unknown.* That’s
* But there’s a somewhat more precise relationship between the normal, χ2 + t distributions.
* Suppose we scale our χ2 data by dividing it by the dF + then take a set of normally distributed variables + divide them by the square root of our scaled χ2 variable
* *We can obtain an F distribution by taking the ratio between 2 scaled χ2 distributions.*