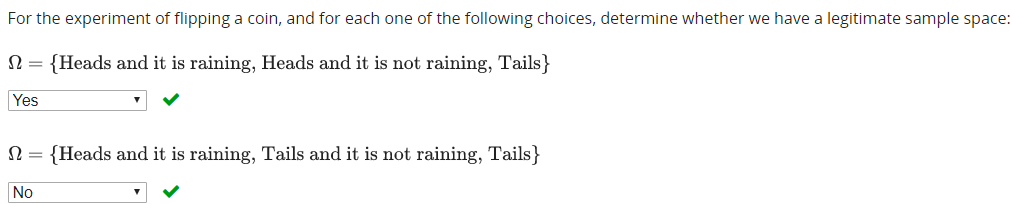
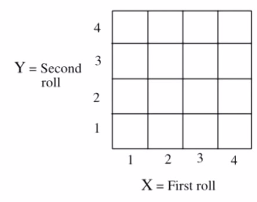
* **Probabilistic** way of thinking involves understanding the nature of probabilistic models, the key concepts, + the mathematical language that goes with them
* Why study probability?
* Until quite recently, **scientific literacy** meant calculus, some physics, + some chemistry.
* W/ the more recent addition of familiarity w/ CPUs + computation, this was all you needed to know in order to make sense of the world.
* But these days, there's not much you can understand about what is going on around you if you don’t understand the **uncertainty** attached to *pretty much every phenomenon.*
* More likely to have to deal w/ uncertainty while analyzing noisy data rather than having to calculate integrals
* **Probability** is now a central component of scientific literacy.
* What is it that has changed and caused this shift?
* 2 main factors.
* As science + engineering move forward, we end up dealing w/ more + more **complex systems**
* In a complex system, we cannot expect to have a *perfect* model of *each* component or to know the *exact* state of *every* piece of the system.
* Uncertainty is now at the foreground + needs to be modeled.
* We live in an **information society**.
* Data + information play an increasingly central role, both in our individual lives + in the economy as a whole.
* Now, data + information are only useful b/c they can tell us something we did not know.
* Their reason for existence is to *reduce uncertainty*.
* But if your goal is to reduce uncertainty, you'd better understand its nature + have the tools to describe it + analyze it
* This is why probability theory + its children-- **statistics** + **inference**— are a must.
* Think of any scientific field, + quickly realize that maybe, other than the motion of the planets, everything else involves uncertainty + calls for probabilistic models.
* Physics + Quantum mechanics has taught us nature is inherently uncertain
* Biological evolution progresses through the accumulation of many random effects, like mutations, w/in an uncertain environment.
* The haystack of biological data we’re accumulating + that needs to be sifted using statistical tools in order to make progress in the biomedical sciences.
* Communications + signal processing 🡺 These fields are, almost by definition, a fight against noise, or an effort to clean signals from the noise nature has added.
* Management 🡺 Customer demand is random, + you want to be able to model + predict it.
* Finance 🡺 Markets are uncertain, + whoever has the best methods to analyze financial data has an advantage.
* Transportation systems 🡺 Random disruptions due to weather or accidents are a major concern.
* Trends in social networks 🡺 spread like epidemics but in ways that are hard to predict
* The message is clear.
* Most phenomena of interest involve significant randomness + the only reason we collect + manipulate data is b/c we want to fight this randomness as much as we can.
* The 1st step in fighting an enemy like randomness is to study + understand your enemy
* **A probabilistic model** = a quantitative description of a situation, phenomenon, or experiment whose *outcome is uncertain*
* A model of a *random* phenomenon/experiment
* Putting together such a model involves 2 key steps.
* 1) Describe the possible outcomes of the experiment by specifying a **sample space**.
* 2) Then, specify a **probability law**, which assigns probabilities to outcomes/collections of outcomes
* The **probability law** tells us, for example, whether 1 outcome is much more likely than some other outcome.
* The **axioms** of probability theory = certain basic properties probabilities have to satisfy in order to be meaningful (ex: cannot be negative)
* There will be very few axioms, but they’re powerful, have lots of consequences, + imply many other properties that were not part of the axioms.
* **Discrete models** are conceptually much easier, while **Continuous models** involve some more sophisticated concepts

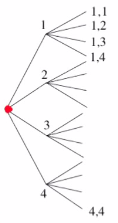
**Sample space**

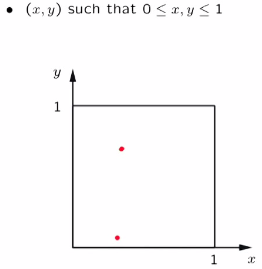
* **A probabilistic model** = A model of a *random* phenomenon/experiment
* Step 1) describe possible outcomes
* Step 2) Describe belief’s about likelihood of the outcomes
* Ex: Coin flip 🡪 2 possible outcomes = Heads, Tails 🡪 sample space
* Designated w/ Omega Ω 🡺 a **set** (unique) of possible outcomes
* **Elements** of a set should be
* **Mutually Exclusive:** if 1 outcome happens, it’s not possible another outcome happens
* **Collectively Exhaustive:** Together, all elements of a set exhaust all possibilities
* To summarize, the set should be such that, at the end of an experiment, you should be always able to point to 1, and *exactly 1*, of the possible outcomes + say “this is the outcome that occurred”
* **Physically different** (different in all relevant aspects but perhaps not in irrelevant aspects) outcomes should be distinguished in the SS + correspond to *distinct* points.
* Suppose you flip a coin + see whether it resulted in H or T 🡪 perfectly legitimate SS of 2 points
* Together these 2 outcomes exhaust all possibilities + 2 outcomes are mutually exclusive, so this is a very legitimate sample space for this experiment.
* Now suppose while you were flipping the coin, you looked outside the window to check the weather
* Now, you could say the SS is really: {H + rain, H + no rain, T + rain, T + no rain}
* This set, consisting of 4 elements, is also a perfectly legitimate SS for an experiment of flipping a coin
* The elements of this SS are mutually exclusive + collectively exhaustive: Exactly 1 of these outcomes is going to be true at the end of the experiment.
* But this SS involves some irrelevant details, so the *preferred* SS for describing the flipping of a coin is the simpler one 🡪 is at the right granularity, given what we're interested in.
* Ultimately, the question of the “right” SS depends on what kind of questions you want to answer.
* Ex: Theory that weather affects behavior of coins 🡪 in order to play w, might want to work w/ the 2nd SS.
* This is a common feature in all of science: *Whenever you put together a model, you need to decide how detailed you want your model to be.*
* The right level of detail = the one that captures aspects that are relevant + of interest to you
* 

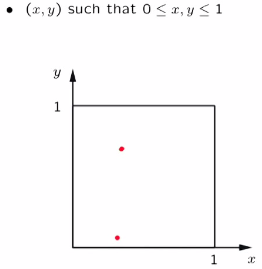
**Sample space examples**

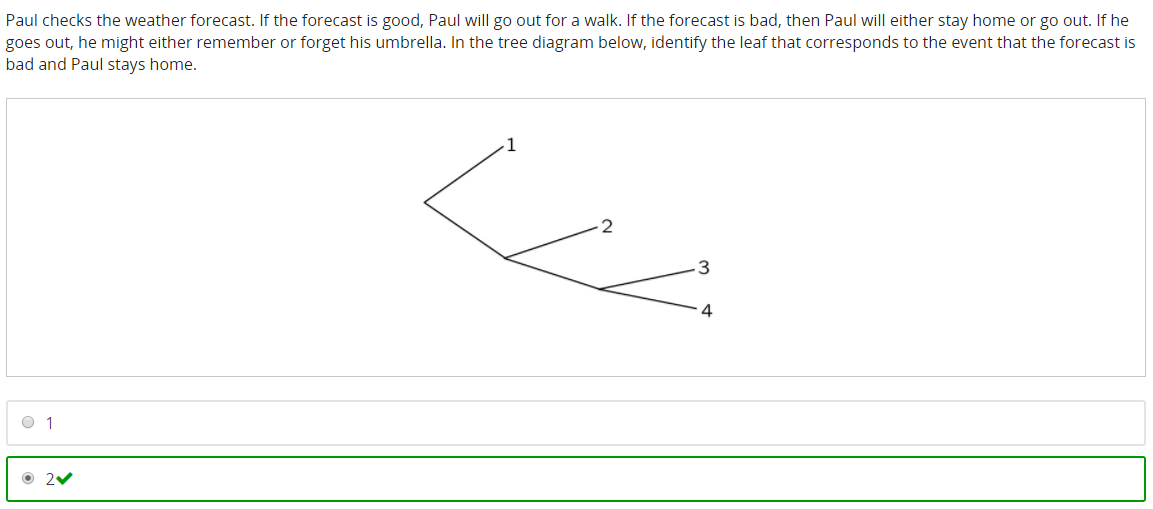
* Sample spaces are **sets** + a set can be **discrete, finite, infinite, continuous**, + so on
* Ex: SS that is discrete + finite 🡪 take a tetrahedral die (4 faces) + roll it once + then roll it again.
* We’re not dealing w/ 2 probabilistic experiments but w/ a *single* probabilistic experiment that involves 2 rolls of the die w/in that experiment.
* 1 possible representation of the SS is:
* Take note of the result of the 1st roll + then take note of the result of the 2nd roll to get a pair of numbers.

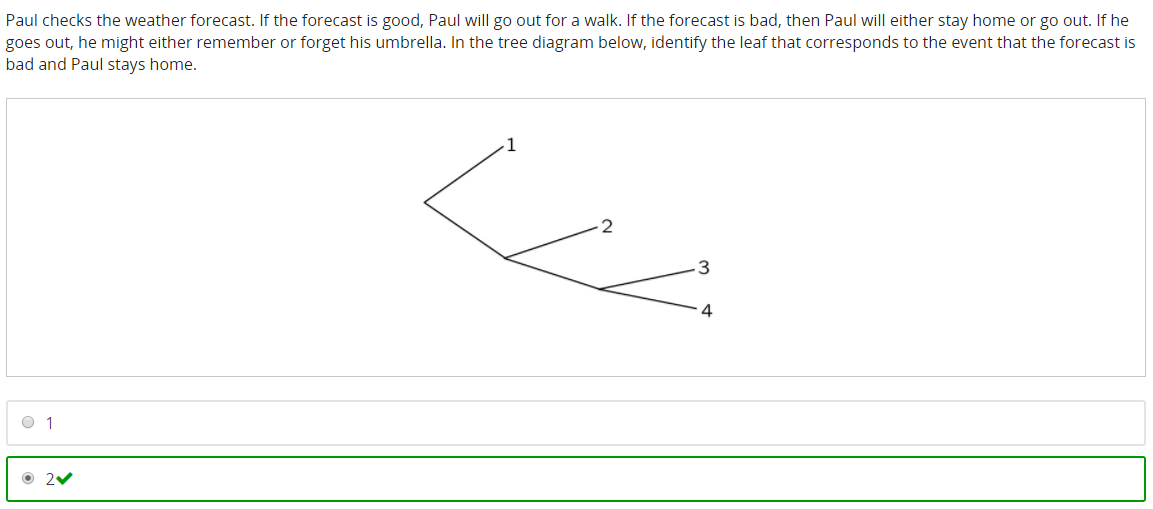
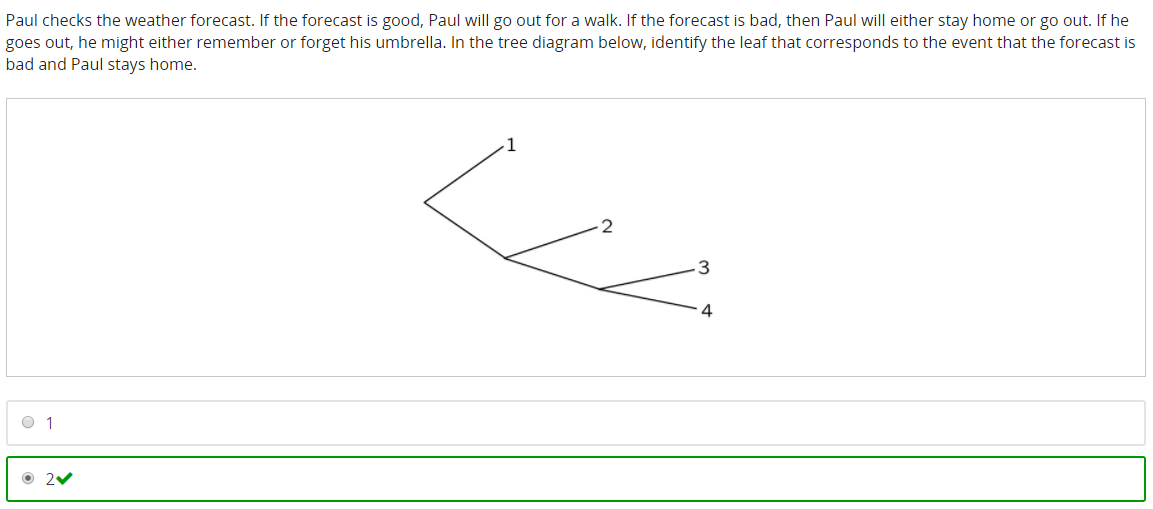


* Notice some outcomes are pretty closely related 🡪 observe a 2 + a 3, but distinguish those 2 outcomes b/c they happen in different order
* The order in which they appear may be of interest to us, + so we make this distinction
* This is a case of a model in which the probabilistic experiment can be described in **phases** or **stages**.
* Think about rolling the die once + then going ahead w/ the 2nd roll = 2 stages.
* A useful way of describing the SS of experiments w/ several stages (either real or imagined stages) is by providing a **sequential description** in terms of a tree.
* 
* 1st stage = **root** + endpoints = **leaves**.
* Experiment starts 🡪 carry out 1st phase = 1st roll + see what happens 🡪 then take note of what happened in the 2nd roll 🡪 then follow the branches that correspond to these observations -🡪 end up at a particular leaf
* Here, we have 16 possible outcomes = discrete + finite.
* SSs can also be infinite + continuous sets
* Ex: Have a rectangular target + you throw a dart on it
* Suppose you are so skilled that no matter what, when you throw the dart, it always falls inside the target.



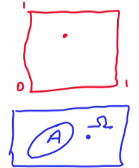


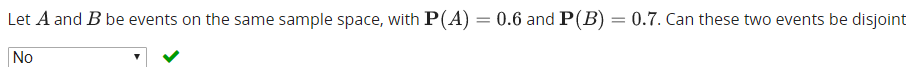
* Once the dart hits the target, you record the coordinates x + y of the particular point that resulted from your dart throw + we record x + y w/ infinite precision.
* So x + y are real numbers + in this experiment, the SS = the set of (x, y) pairs that lie between 0 + 1
* 



**Probability axioms**

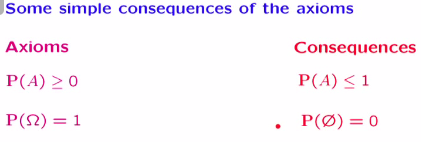
* The 2nd + much more interesting step in constructing a probabilistic model is to specify which outcomes are more likely to occur + which ones are less likely by assigning probabilities to the different outcomes.
* However, as we try to do this assignment, we run into difficulty, which is the following.
* Ex: Experiment involving a continuous sample space 🡪 probability of a particular point would essentially be 0 (Hitting the center of a dartboard exactly w/ infinite precision should be 0)
* So it's natural that in such a continuous model, any individual point should have a 0 probability.
* For this reason, instead of assigning probabilities to individual points, we instead *assign probabilities to whole* **subsets** of the SS = an **event**
* See subset A w/in Ω.

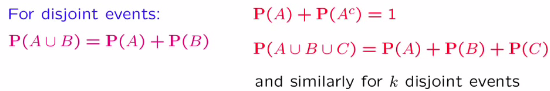


* We're going to assign a probability to that subset A, which we're going to denote P(A)
* Probabilities will be assigned to subsets = will not cause difficulties in a continuous case, b/c even though individual points have 0 probability, the odds a dart falls in the upper half should be a reasonable positive number.
* So even though individual outcomes may have 0 probabilities, *sets of outcomes generally would be expected to have positive probabilities.*
* So we assign probabilities to various subsets of a SS = **events**
* Once we carry out an experiment + observe the outcome of it, we say “event A has occurred”, or if the outcome falls outside the set A, we say “event A did not occur”
* The rules of probabilistic models = the rules probabilities should satisfy.
* Shouldn't be completely arbitrary.
* Always given in the range between 0 + 1
* These rules any probabilistic model should satisfy = **the** **axioms of probability theory**
* Axiom 1) Non-negativity 🡪 probabilities will always be non-negative numbers
* Axiom 2) U🡪 probability of Ω (whole set) should always be equal to 1
* “We have absolute certainty that event Ω is going to occur” 🡪 some event w/in Ω *will* occur
* Capture this certainty by saying the probability of event Ω, P(Ω), is equal to 1.
* Axiom 3) Additivity
* 
* Quick reminder about set theoretic notation. If we have two sets, A + B set:
*  = "A **intersection** B" = the collection of elements that belong to both A + B.
*  = "A **union** B" = the set of elements that belong to A OR B OR both. So in terms of this picture, the union of the two sets would be this blue set.
* So the 3rd axiom says “if we have 2 sets/events/subsets of the SS that are **disjoint** (their intersection has no elements = an empty set = ), the probability the outcome of the experiments falls in the **union** of A + B == the sum of the probabilities of the 2 individual sets P(A) + P(B)
* This is called the additivity axiom.
* So it says we can add probabilities of different sets when those 2 sets are disjoint.
* In some sense, we can think of probability as being 1 lb. of some substance which is spread over our SS + P(A) is how much of that substance is sitting on top of a set A.
* Therefore, the 3rd axiom is saying is the total amount of that substance sitting on top of A + B is how much is sitting on top of A + how much is sitting on top of B in the case whenever the sets A + B are disjoint from each other.
* Other than needing to refine the additivity axiom a bit, these 3 axioms are the only requirements in order to have a legitimate probability model.
* Shouldn't we, for example, say probabilities cannot be greater than 1?
* We do not want probabilities to be larger than 1, but we do not need to say it
* Such a requirement follows from what we have already said + the same is true for several other natural properties of probabilities.
* 
* Have to have some common elements b/c they add up to more that 1.0

**Simple properties of probabilities**

* What you might think of as “missing axioms” are actually implied by the axioms already in place.

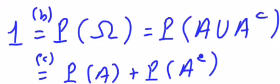




* Consider a SS, a subset A, and its **complement** A(c) = set of all elements that do not belong to set A.
* A set together w/ its complement make up the entire SS, Ω  
  Also, if an element belongs to a set A, it does not belong to its complement 🡪 the intersection of a set w/ its complement = the empty set Φ
* Now argue as follows: We have “the probability of the entire SS, P(Ω), = 1”
* The SS can be written as the **union** of an event + its complement



* Since a set + its complement are **disjoint**, we can apply the **additivity axiom**



* P(intersection) = P(A) + P(A(c))
* Based on this relation, we can also write that P(A) = 1 – P(A(c)), + b/c, by the non-negativity axiom, this result is non-negative 🡪 1 - something non-negative <= 1.



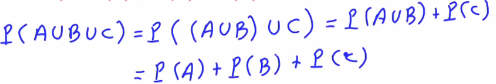
* Using here the non-negativity axiom, we established another property 🡺 Probabilities are always <= 1.
* Finally, let us note that 1 = P(Ω) + P(Ω(c)) = P() of a set + P() of the complement of that set
* Remember P(Ω) = the probability of the entire SS, which itself = 1
* The complement of the entire SS consists of all elements that do NOT belong to the SS, but since the SS contains ALL possible elements, *its complement is just the empty set*
* From this relation we get the implication the probability of the empty set P(Φ) = 0.
* This establishes yet 1 more of the properties that we had just claimed a little earlier.
* Proof of the generalization of our additivity axiom from the case of 3 disjoint events



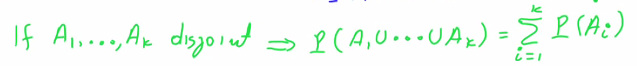
* So we have our SS w/ 3 events/subsets that are disjoint in the sense that any 2 of those subsets have no elements in common
* We're interested in the probability of the union of A, B, + C
* The additivity axiom applies to the case of the union of 2 disjoint sets, while here we have 3
* We can do the following trick 🡪 think of the union of A, B, + C as consisting of the union of A and B with set C.
* Formally, what we're doing is expressing the union of these 3 sets as follows.
* Form 1 set by taking the union of A w/ B + the overall union can be thought of as the union of these 2 sets w/ the other set C.



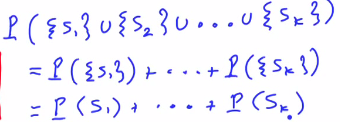
* Since the three sets are disjoint, this implies that the 1st set is disjoint from the 2nd, so we can use the additivity axiom to write this overall probability as P(A U B) + P(C).
* Then since sets A + B are disjoint, write the 1st term as P(A) + P(B) to end up w/ P(A) + P(B) + P(C)



* Can follow this line of proof to write an argument for the case of 4 events + so on
* Can that if sets A1-Ak are disjoint, the probability of the union of those sets = the sum of their individual probabilities.

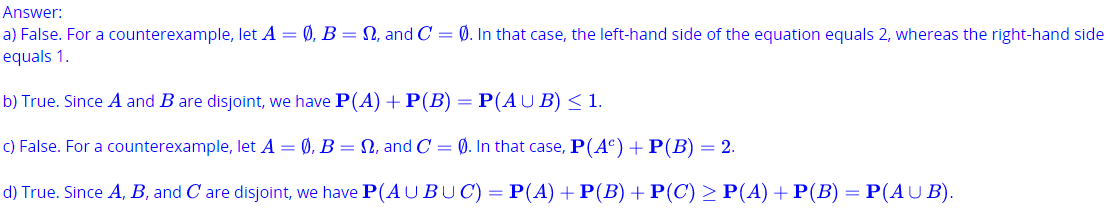


* This is the generalization to the case when dealing w/ the union of *finitely* many *disjoint* events.
* A very useful application of this = the case where we want to calculate the probability of a finite set.
* An SS w/ some particular elements S1-Sk 🡪 these elements together form a finite set.
* Think of this finite set that consists of k elements as the union of several little sets that contain 1 element each.
* Theoretically what we're doing is taking a set w/ k elements + writing it as the union of a set that contains just element S1, a set that contains just the element S2, + so on, up to the k-th element
* Assuming, of course, that these elements are all different from each other.
* In this case, these single element sets are all disjoint.
* Using the additivity property for a union of k disjoint sets, we can write this as the sum of the probabilities of the single element sets.



* ` 
* For each one of the following statements, determine whether it is true or false. *Note:* "False" means "not guaranteed to be true."



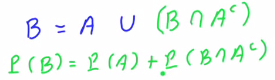


**More Properties of Probabilities**

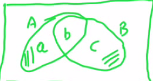
* There are additional properties of probability laws which are, again, consequences of the 3 axioms
* 1) 
* If we have 2 sets + 1 is smaller than the other (SS(b) > SS(a)), the probability an outcome falls inside B should be at *least as big* as the probability the outcome falls inside A.
* 
* How do we prove this formally?
* Set B can be expressed as a **union** of 2 pieces 🡪 set A + all elements of B that do not belong in A



* These belong to the **complement** of A 🡪 express set B as the union of 2 disjoint pieces.
* Therefore we can apply the additivity axiom + write P(B) = P(A) + P(B N A(c))



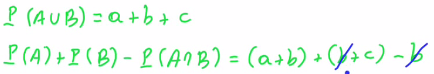
* Since probabilities are non-negative, the 2nd term is at least as large as P(A) (is greater than or equal to P(A)), therefore P(A) <= P(B)
* 2) 
* i.e. the probability of the union of 2 sets where the 2 sets are *not necessarily* disjoint.

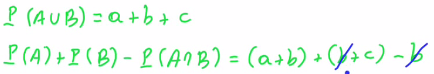


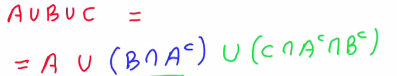
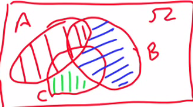
* Union of A + B consists of 3 pieces 🡪 elements of A that do not belong to B (A intersect B(c)), the **intersection** of A + B, + elements in B that do not belong in A (B intersect A(c))

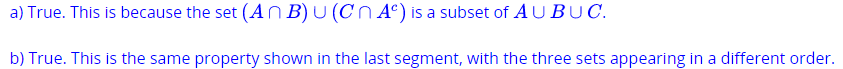


* So P(A union B) consists of these 3 pieces
* By the additivity axiom, the P(A U B) = the sum of the probabilities of these 3 pieces.



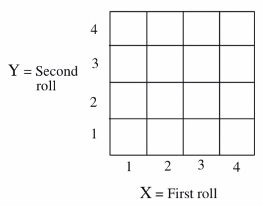
* P(A) + P(B) – P(A intersect B) =
* Set A consists of 2 pieces w/ probabilities a + b.
* Set B consists of 2 pieces w/ probabilities b + c.
* Must then subtract the probability of the *intersection* 🡪 b
* This cancels out 1 b + we are left w/ a+ b + c
* 
* 1 particular consequence of this equality derived is:
* Since P(A interest B) is always non-negative, P(A U B) is always <= P(A) + P(B)
* 
* This inequality here is quite useful whenever we want to argue a certain probability is smaller than something
* This property is called the **union bound**.
* 3) 
* Derive an expression/way of calculating the probability of the union of *three* sets that’re NOT necessarily disjoint.
* W/in the SS we have 3 sets/events
* We are going to use a set theoretic relation + will express the union of these 3 sets as the union of 3 disjoint pieces.
* 1 = Set A itself
* 2 = Part of B which is outside A 🡪 intersection of B w/ complement of A
* 3 = whatever is left in order to form the union of the 3 sets = that part of C that does not belong to A nor belong to B = C intersection with A complement + B complement.



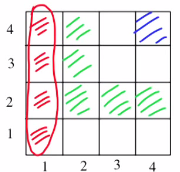
* Intersection of 2 sets is the same no matter in the order
* Notice these 3 pieces are disjoint from each other, so by the additivity axiom, we get the 3rd consequence above
* 
* For each one of the following statements, determine whether it is true or false. *Note:* “False" means “not guaranteed to be true."
* 
* 
* 

**Discrete SS Example**

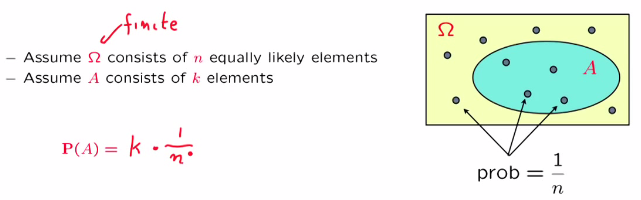
* 2 rolls of tetrahedral die = 16 possible outcomes



* Assign probability laws w/ assumption each outcome has 1/16 chance of happening (equally likely)
* **P(X = 1)** 🡪 can happen in 4 ways (4 different Y values) = 4/16 = ¼
* Let Z = min(X, Y) 🡪 the smaller of the 2 numbers from the 2 rolls
* Now calculate P(Z = **4)** 🡪 find all outcomes where 4 is the smaller number 🡪 only “happens” when BOTH outcomes are 4 (neither can be > 4, so it’s the min) 🡪 1/16
* **P(Z = 2)** 🡪 all outcomes w/ Y = 2 and X >= 2 🡪 X = 2, 3, or 4, then all outcomes X = 2 w/ Y >= 2 🡪 Y = 2, 3, or 4 🡪 6/16 – 1/15 b/c X = 2 and Y = 2 appear in both sets 🡪 in green below

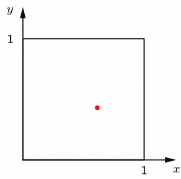


* P(value in 1st roll is strictly larger than value in the 2nd roll) 🡺 {21, 32, 31, 43, 42, 41} = 6/16
* P(sum of the values obtained in the two rolls is an even number) 🡪 {11, 13, 22, 24, 31, 33, 42, 44} = 8/16 = 1/2
* This particular example is a special case of **a Discrete Uniform Law.**
* In a **Discrete Uniform Law**, we have a *finite* SS w/ n equally-likely (assumed) elements.
* P(Ω) = 1, so this means that each element must have P(1/n)
* That's the only way the sum of the probabilities of the different outcomes would be equal to 1 as required by the normalization axiom.
* Consider now some subset of the SS, event A w/ exactly k elements.
* P(A) is the sum of probabilities of *its* k elements, each of whom have probability of 1/n.
* We can find the P(set A) when we have a discrete uniform probability law 🡪 can calculate probabilities by simply counting the number of elements of Ω (= n) + counting the number of elements of the set A (k)

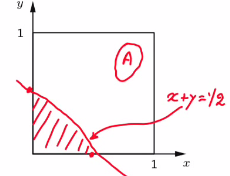


**Continuous SS Example**

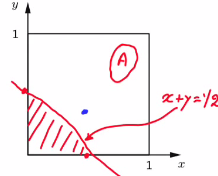
* Ex: Throwing dart into unit square target, guaranteed the dart falls somewhere in the set
* So SS = unit square itself

* Not given a probability law 🡪 must derive
* Assume our probability law is a **uniform probability law** = *probability of any particular subset of the SS = area of that subset*
* This is an arbitrary choice of probability law (nothing in our assumptions so far forces us to make this choice)
* Find P(the sum of the 2 numbers of the dart <= 1/2):
* 
* Useful to work in terms of a picture:

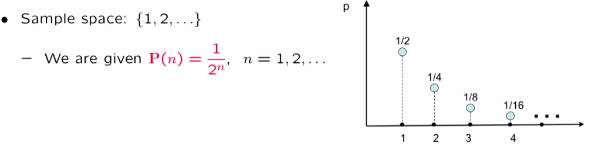


* Anything below the red line satisfies this requirement
* Area = triangle = 1/2Bh = ½(1/2)(1/2) = ½(1/4) = **1/8**
* Find P(event of only a single element/set of only a single point) = P({0.5, 0.3})



* Probability of any particular subset = area of that subset, and area of a single point = 0
* From these examples, abstract the following to calculate probability in 4 steps:
* Specify SS
* given some word description of a problem/probabilistic experiment
* Specify probability law
* any you like, but for results to be useful, good if it capture the real-world phenomenon we’re trying to model
* ID event of interest
* Might be described in a loose manner so we must describe it mathematically
* If possible, always good to describe in a picture
* Calculate P(event of interest)
* In principle, a probability law specifies the probability of *every* event, + there's nothing else to do.
* But quite often the probability law will be given in some *implicit* manner, for example, by specifying the probabilities of only *some* events.
* In that case, you may have to do some additional work to find the probability of the particular event you care about.
* This will sometimes will be easy, sometimes it may be complicated.
* In either case, by following this 4-step procedure + by being systematic you will always be able to come up with a single correct answer.
* Consider a SS that is the rectangular region [0,1]×[0,2], i.e., set of all pairs (x, y) that satisfy 0 <= x <= 1 + 0 <= y <= 2. Consider a **uniform probability law**, under which the probability of an event = 1/2 of the area of the event. Find the probability of the following events:
* The 2 components x and y have the same values 🡪 creates straight diagonal linear line from origin = **0**
* The value, x, of the 1st component is >= the value, y, of the 2nd component 🡪 area under the diagonal line from above 🡪 forms triangle = ½(1)(1) = ½ then 1/2\*1/2 = **¼**
* The value of x^2 is larger than or equal to the value of y 🡪 corresponds to the region below the curve y = x^2, where x ranges from 0 to 1.
* The area of this region is the integral of x^2 from 0-1 🡪 x^3/3 🡪 1^3/3 – 0^3/3 = 1/3
* Then halve it due to our law = **1/6**

**Countable Additivity**

* Now for an example with an INFINITE discrete SS
* Ex: experiment w/ outcome = arbitrary positive Int. 🡪 keep tossing coin + outcome = 1st time we see H
* Any possible int. (toss #) is possible, so our SS is infinite
* Specify probability law 🡪 *remember a probability law should determine P() of EVERY event/EVERY subset of the SS* 🡪 i.e. the P() of EVERY set of positive integers
* Instead see the P() of a events w/ a single element:
* 
* i.e. P() of observing n = 1 / 2^n
* If this info enough to determine P() of any subset?
* To find out, first do a sanity check to see if these given #’s seem like legit probabilities
* Do they add to 1 🡪 sum over all possible values of n (infinite sum)

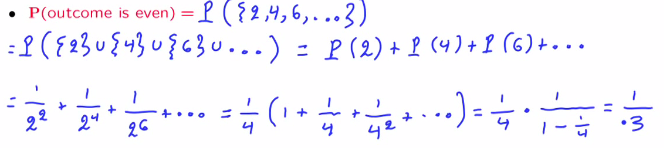
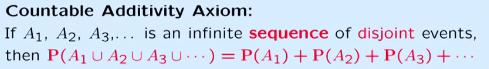


* Factor out 1/2, which lowers the exponent from n to n – 1, which is the same as making out starting n for the sum equal to n = 0 rather than n = 1



* This is ½ multiplied by an **infinite geometrics series,** which has a formula = 1 over 1 minus the number whose power we’re taking (1/2), which ends up being 1/2 – 1/(1/2) = 1



* It seems we have the basic element to have a legitimate probability law
* How do we calculate the P() of some general event?
* Ex: P(outcome is even)
* This is P(infinite set consisting of all even integers) 🡪 P({2, 4, 6,…})
* Can write this as the union of single element sets 🡪 = P({2} U {4} U {6} …)
* Notice we now have a union of sets that are *disjoint* (mutually exclusive) 🡪 can use additive property 🡪 P(2) + P(4) + P(6) …. = 1/2^2 + 1/2^4 + 1/2^6 …
* Factor out ¼ to be left with ¼\*(1 + 1/2^2 + 1/2^4) 🡺 ¼\*(1 + ¼ + 1/4^2 + …) 🡺 ¼ \* infinite some of a geometric series = ¼ \* (1 / (1 – ¼)) = **1/3**
* 
* But is this calculation correct?
* We used the additivity property, but this only talks about disjoint events of *finite* amounts of subsets
* So that step is actually *not allowed*
* But we want our theory to allow this kind of calculation, so to get out of this dilemma, introduce a new axiom to do so
* 
* This strengthens the *finite* additivity axiom
* **Sequence** = events can be arranged such that we can talk about the 1st event, the 2nd event, + so on
* To appreciate the issue that arises here + to see why the word **sequence** is so important, let consider the following calculation:
* SS = unit square, consider model where P(set) = set’s area
* Look at P(Ω), which = area of unit square = 1 🡪 can be thought of as the union of various sets of single points (1 element each) = P(U{(x,y)}) over all points in the unit square

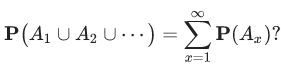


* Subsets are disjoint here, so different points = disjoint single element sets, so the additivity axiom tells us the P(this union) = sum of P()’s of all single-element subset in the set
* We know single-element subsets have P() = 0, so this sum = 0



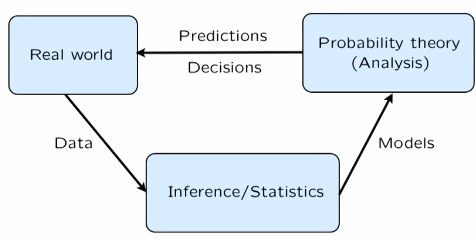
* On the other hand, by our probability axioms, P(Ω) = 1, but this is saying P(Ω) = 0
* The catch = there is nothing in the axioms introduced so far/the properties have established that would justify this step, so this step here is questionable:



* Might argue that the unit square = the union of disjoint, single-element sets, which is the case in additivity axioms.
* *But the additivity axiom only applies when we have a SEQUENCE of events,* + this is NOT what we have here.
* “Additivity holds only for countable sequences of events”
* The above is NOT a union of a SEQUENCE of single element sets.
* In fact, there is no way the elements of the unit square can be arranged in a sequence.
* The unit square is said to be an **uncountable set**, a deep and fundamental mathematical fact that essentially says is there are 2 kinds of infinite sets.
* **Discrete (countable) sets** = whose elements can be arranged in a sequence, like integers.
* **Uncountable sets** (such as the unit square or the real line) = whose elements cannot be arranged in a sequence.
* After all these discussion, you may now have legitimate suspicions about the models we’ve been looking at.
* Is area a legitimate probability law? Does it even satisfy countable additivity?
* This question takes us into deep waters + has to do w/ a deep subfield of mathematics, **Measure Theory**.
* Fortunately, it turns out that all is well + area IS a legitimate probability law + DOES indeed satisfy the countable additivity axiom, *as long as we only deal w/ nice subsets of the unit square.*
* Fortunately, the subsets that arise in whatever we do in this course will be "nice".
* Subsets that are NOT nice are quite pathological + we will not encounter them.
* At this stage we are not in a position to say anything more that would be meaningful about these issues b/c they're quite complicated + mathematically deep.
* We can only say that there are some serious mathematical subtleties.
* But fortunately, they can all be overcome in a rigorous manner + for the rest of this class, you can just forget about these subtle issues.
* Let the sample space be the set of positive integers + suppose P(n) = 1/2^n, for n = 1, 2, …. Find the probability of the set {3,6, 9,…}, that is, of the set of positive integers that are multiples of 3.
* **1/2^3 + 1/2^6 + 1/2^9 … 🡪 1/8\*(1 + 1/8^2 + 1/8^3) 🡪 1/8 – (1/(1 – 1/8)) = 0.14285714285**
* Let the sample space be the set of positive integers. Is it possible to have a “uniform" probability law, that is, a probability law that assigns the same probability c to each positive integer?
* **No. Suppose c = 0. Then, by countable 1 = P(Ω) = P({1} ∪ {2} ∪ {3}, …) = P({1}) + P({2}) + P({3}) + … = 0 + 0 + 0+⋯=0, which is a contradiction.**
* Suppose c > 0. Then, there exists an integer k such that kc > 1. By additivity, P({1, 2, …, k}) = kc > 1, which contradicts the normalization axiom.
* If there is a uniform probability distribution on the set of positive integers, then every integer must have the same probability = c
* 2 possibilities: c = 0, or c > 0. This shows that both of these possibilities lead to a contradiction.
* Therefore neither can be true, which means our original assumption (there is a uniform probability distribution on the set of positive integers) cannot be true either.
* Let the sample space be the 2D plane. For any real number x, let Ax be the subset of the plane that consists of all points of the vertical line through the point (x, 0) 🡪 i.e., Ax = {(x, y) : y ∈ ℜ}
* Do the axioms of probability theory imply that the probability of the union of the sets Ax (which is the whole plane) is equal to the sum of the probabilities P(Ax)?
* **NO b/c a set of real numbers y ∈ ℜ in the collection of sets Ax is NOT countable (i.e. cannot be arranged in a sequence) so the additivity axiom does not apply**
* Do the axioms of probability theory imply that 
* (In other words, we consider only those lines for which the x coordinate is a positive integer.)
* **YES b/c the countable additivity axiom applies b/c we’re dealing w/ a sequence (countable collection) of disjoint events**

**Interpretations and uses of probabilities**

* Most narrow view 🡪 **Probability Theory** = a branch of math in which we start w/ some **axioms** + consider **models** that satisfy these axioms, establish some **consequences**, which are the **theorems**
* Can do all that without ever asking the question of what the word "probability" really means.
* Yet, 1 of the theorems of probability theory is that probabilities can be interpreted as **frequencies**, very loosely speaking.
* If I have a fair coin, and I toss it infinitely many times, the fraction of H I will observe will be 1/2
* In this sense P(A) can be interpreted as the frequency w/ which event A will occur in an infinite number of repetitions of the experiment.
* But is this all there is?
* If we're dealing w/ coin tosses, it makes sense to think of probabilities as **frequencies**.
* But consider a statement such as the "current president of my country will be reelected in the next election with probability = 0.7".
* It's hard to think of this number, 0.7, as a frequency. It does not make sense to think of infinitely many repetitions of the next election.
* In cases like this, and in many others, it is better to think of probabilities as just some way of **describing our beliefs**.
* And if you're a betting person, probabilities can be thought of as some numerical guidance into what kinds of bets you might be willing to make.
* But if we think of probabilities as beliefs, you can run into the argument that beliefs are subjective
* Isn't probability theory supposed to be an objective part of math + science? Is probability theory just an exercise in subjectivity?
* Well, not quite. There's more to it.
* **Probability**, at the minimum, gives us some rules for thinking systematically about uncertain situations.
* And if our probability model (subjective beliefs) have some relation w/ the real world, probability theory can be a very useful tool for making predictions + decisions that apply to the real world.
* Now, whether your predictions + decisions will be any good will depend on whether you have chosen a model that provides a good-enough representation of the real world
* How do you make sure that this is the case?
* There's a whole field, the field of statistics whose purpose is to complement probability theory by using data to come up w/ good models.



* Real world *generates data* 🡪 Field of statistics + inference *uses these data to come up w/ probabilistic models* 🡪 Once we have a probabilistic model, we use probability theory + the analysis tools the model provides to us.
* The results that we get from this analysis lead to predictions + decisions about the real world.