Probability and Inference

Christopher Lee

Department of Chemistry and Biochemistry, UCLA

Announcements

- let's quickly finish defining a few more bio terms.
- This week is **mission training only** ("Homework 1", and no project). Please complete it before Friday's discussion section. See the link in CCLE Week 1.
- new reading assignment for today's lecture will be posted today.
- today: basic probability review, and Bayes Law, covering everything needed for homework 1.
- Note: lecture slides posted before each topic, in case you want to download them before class.

Probability Concepts Review

- a solid understanding of probability is required for this class (Stats 100A prerequisite).
- but regardless, this always comes up as something people seem to need some review for, before we dive deeper into probabilistic modeling.
- So we're going to spend several classes going over essential probability concepts, e.g. conditional probability and the chain rule.
- we'll work through many concept test questions in class, to help you sort out anything that confuses you, and become adept at using these simple but powerful tools.

Your Mission This Week: Is Mrs. X a Disease Carrier?

- Mrs. X comes from a genetic background that has an unusually high rate of a particular X-chromosome linked genetic disease.
- Mrs. X does not show disease symptoms, but she still might be a *disease carrier* (i.e. has a copy of the disease mutation that she could pass on to her children, especially sons, who only receive an X chromosome from their mother, not their father).
- If Mrs. X has one (or more) sons *without* disease symptoms, how exactly would that change the risk that she is a disease carrier?

Your Mission: Mrs. X wants you to compute her exact risk of being a disease carrier, based on these family data.

Basic Mendelian Genetics (Plants, Animals, etc.)

- *Mendelian inheritance*: the standard pattern of gene transmission in plants and animals, in which each individual has two copies of a gene (one copy from each of its two parents), and passes on one of its copies (chosen at random) to each of its children.
- *recessive mutation*: Because of this, most mutations that damage a gene function are "recessive", which means that *both* copies of the gene must have a damaging mutation, in order to produce the mutation's phenotype (i.e. disease symptoms). An individual with one mutated copy plus one normal copy would simply have a normal phenotype (i.e. no disease symptoms).
- *dominant mutation*: A mutation that can cause its phenotype even if it is present on only one of the two copies is said to be dominant.

Mendelian Inheritance

- Say we have a "normal" gene (+), and a mutation that knocks out that gene's function (-).
- A person with one normal copy and one mutant copy (+-) will show no symptoms (the mutation is "recessive").
- If two such people have a child, the child gets one copy of the gene (chosen at random) from each parent, so the probabilities are 25% ++, 50% +-, 25% --.
- Exactly the same as flipping two coins...
- Only if the child is -- will he show symptoms.

Sex Chromosomes & the Disease of Kings

- Women have two copies of the X chromosome (XX), whereas men have one copy of X and one copy of the Y chromosome (XY).
- Each child randomly receives one of the two chromosomes from each parent, i.e. one X from mom and either X or Y from dad, giving either XX (girl) or XY (boy).
- So if mom is X^+X^-, her son has 50% chance of being X^-Y
- Disease example: hemophilia (blood fails to clot) is a recessive disease mutation on the X chromosome, so it is far more common (actual disease symptoms) among men than women. And historically it was associated with European royal families (due to inbreeding)!

Completing This Week's Mission

- Mrs. X wants her answer by Friday.
- Go to CCLE Week 1 and click the link for HW1, answer the questions online.
- To get full credit, must answer before Friday's discussion section (noon).
- This is training, not testing: as you answer each part, get immediate feedback that helps you learn from any mistakes you make, and get full credit as long as you really explain your answer.
- Lots of opportunities to go through these lessons in detail in Friday's discussion section.

Mini-language: Hidden vs. Observable Variables

- An **observable** is something you know with zero uncertainty. Concretely, this means a measurement taken directly from some experiment or measurement process you have defined. "What you know".
- A *hidden variable* is any value about which you have some uncertainty. By definition, anything that *isn't* observable. Typically I will symbolize a hidden variable with a Greek letter e.g. θ. "What you want to know".

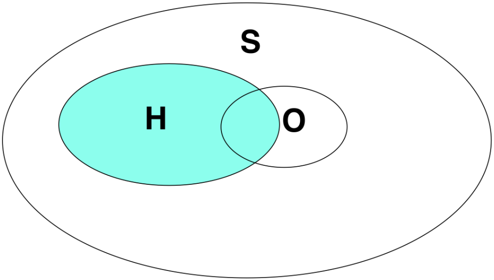
Example:

- observable: measurement in "feet"
- hidden: the true length of some object.

Statistical inference

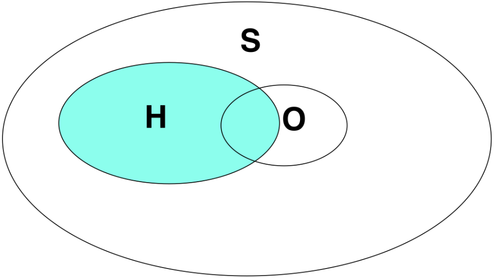
- Observables are *what you know*.
- Hidden variables are *what you want to know*.
- *Statistical inference* is the mathematical process of computing probabilities of hidden variable(s) H based on actual observed values of observable variable(s) O.
- Concretely this means computing $p(H|O)$.

Unconditional probability



$$p(H) = \frac{|H \cap S|}{|S|}$$

Conditional probability



$$p(H|O) = \frac{|H \cap O \cap S|}{|O \cap S|}$$

Let's try some more concept tests

- Connect to wifi: **UCLA_WIFI** (UCLA password) or **UCLA_WEB**
- **https://courselets.org/ct/** gives **Live Session: Join** button.
- Bookmark this for easy access in future classes!
- **Make sure to Join Live Session** (rather than clicking on something else).
- Be careful not to click this course (**UCLA Bioinfo 260A**) on the Courselets.org Public Courses page, or you'll get enrolled a second time!

(If you aren't logged in to Courselets, you can do that by logging in to **ccle.ucla.edu**, click on our class site, go to **Week 1**, and click link to **our problem-solving exercise platform (courselets)**).

Disease Test Reliability

A biotech company has developed a new test for a rare disease (found in less than 1% of the population), which predicts either that a patient has or does not have the disease. The company reports that in a random patient sample the test was 97% accurate (i.e. gave a negative test result) among patients who did not have the disease, and 95% accurate (positive test result) among patients who actually had the disease.

Say you're a physician advising a patient who has just received his result from this test. The patient asks "How reliable is this result?" What specific information would you give him and why?

Disease Test Reliability Answer

The key thing to realize is that what matters for a patient is the conditional probability $p(D|T)$. I.e. given the observation (the test result **T**), what is the reliability vs. uncertainty in forecasting the hidden variable (whether the patient has disease, **D**)?

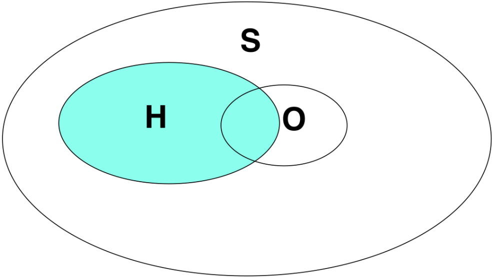
- Note that the question gave you the opposite conditional probabilities $p(T|D)$. These are not relevant to a patient or doctor because they do not go from "what you know" (T) to "what you want to know" (D).
- Estimating $p(D^+|T^+)$ follows straight from the stated numbers: $p(D^+, T^+) < 1\%$, and $p(D^-, T^+) = 3\%$, so $p(D^+|T^+) < 25\%$. Not very reliable! So a positive test result definitely needs to be retested by another method.

This problem of high false positive rate (because the actual disease is rare) is a very common problem in bioinformatics, where our calculations must "scale", e.g. to search for a single disease gene out of the entire genome of 25,000 genes.

Disease Test Reliability Errors

- Many people didn't consider the *direction* of the conditional probability, even though the question's phrasing and answers encouraged you to do that. (The question gave you $p(O|H)$ but asked you about $p(H|O)$). Implies they didn't realize that any conditional probability has two possible directions.
- In particular, people often forget to ask themselves which direction is relevant in real life, i.e. which variable is *hidden* vs. *observable*. Suggestion: remember we can only make inferences (calculate probabilities) of things we want to know (*hidden*) based on things we know (observable). Etch into your minds: *Which variable is hidden? Which varible is observed? Which direction of conditional probability am I being asked for?*
- Etch into your minds: if a (hidden) state is rare, be very worried about the number of false positives (no matter how good the test is)!!
- Some people chased red herrings like "does *reliable* mean 95%? 97%?"

Joint probability

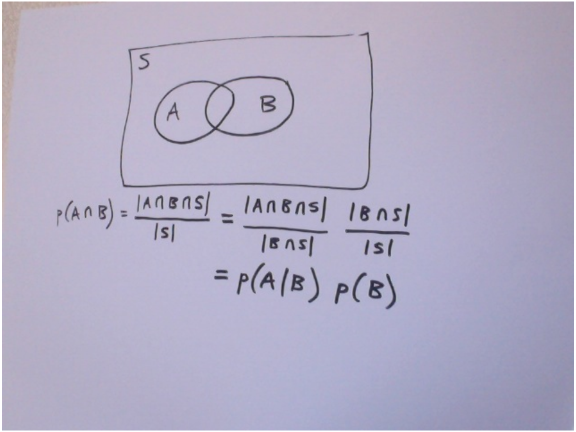


$$p(H \cap O) = \frac{|H \cap O \cap S|}{|S|}$$

Venn Diagram Conditional Probability

Draw a Venn diagram for two intersecting events A and B, and define the conditional probability $p(A|B)$ in terms of regions on the Venn diagram. Then use the Venn diagram to find the mathematical relation between $p(A \cap B)$ versus $p(A|B)$ and $p(B)$.

Venn Diagram Conditional Probability Answer



$$p(A \cap B) = \frac{|A \cap B \cap S|}{|S|} = \frac{|A \cap B \cap S|}{|B \cap S|} \frac{|B \cap S|}{|S|} = p(A|B)p(B)$$

Deriving the Chain Rule

For any joint probability $p(X, Y, Z, W...)$ we can multiply both its numerator and denominator by the probability measure $|Z, W...|$ of some subset of the variables:

$$p(X, Y, Z, W...) = \frac{|X, Y, Z, W...|}{|S|} = \frac{|X, Y, Z, W...|}{|Z, W...|} \frac{|Z, W...|}{|S|}$$

where S is the set of all possible events. But by the definition of probability, these two ratios are just the conditional and unconditional probabilities

$$= p(X, Y|Z, W...)p(Z, W...)$$

Bayes' Law

- Since by symmetry

$$p(H|O)p(O) = p(H \cap O) = p(O|H)p(H),$$

we get the identity:

$$p(H|O) = \frac{p(H \cap O)}{p(O)} = \frac{p(O|H)p(H)}{\sum_h p(O|H=h)p(H=h)}$$

- Bayes' Law is the foundation for many forms of statistical inference.

The Monty Hall Problem

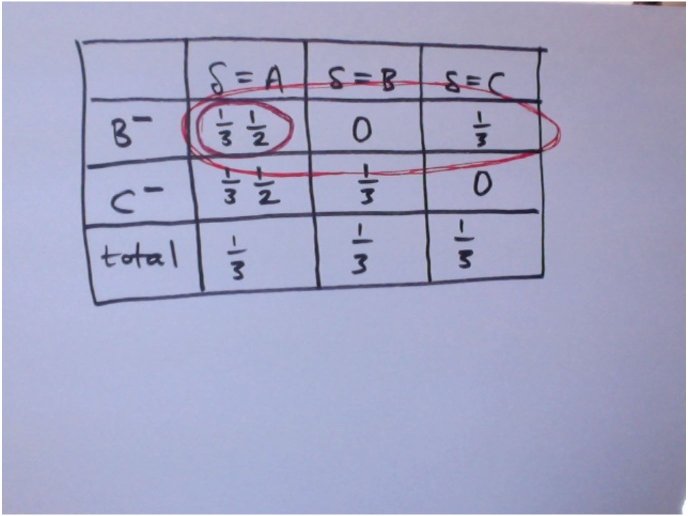
Imagine you're a contestant on the Monty Hall Show. There are three doors, one concealing a valuable prize, and if you pick the right door, you win. But there's a twist: after you say which door you've picked, Monty **always** opens another door and shows you it's **empty** (no prize). Then he asks you whether you want to **switch** your choice to the other closed door. Should you? Does it make any difference?

A Graphical Version of Monty Hall

Draw a table of the joint probabilities of the hidden vs. observed variable for Monty Hall, in which the vertical columns are the three possible hidden states ($\delta = A$, $\delta = B$, $\delta = C$) and the horizontal rows are the two possible observed states (we assume that you initially picked door A, so they are B^- and C^-).

- In each cell in the table write the joint probability of that specific pair of states.
- Finally, circle the two regions of the table that constitute the Venn diagram subset and superset whose ratio defines the conditional probability $p(\delta = A|B^-)$.

A Graphical Version of Monty Hall Answer



	$S=A$	$S=B$	$S=C$
B^-	$\frac{1}{3}\ \frac{1}{2}$	0	$\frac{1}{3}$
C^-	$\frac{1}{3}\ \frac{1}{2}$	$\frac{1}{3}$	0
total	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{3}$

A Graphical Version of Monty Hall Errors

- Some people wrote conditional probabilities, rather than joint probabilities $p(\delta, obs)$ as the question very explicitly asked. I get the feeling that some people are only comfortable with probability of a single variable, not of two or more variables, so their brain gravitated toward writing $p(X|Y)$ instead of $p(X,Y)$. Suggestion: write out simple joint probability tables for two coin tosses (or two dice rolls etc.), and make *sure* you're comfortable with joint probability!
- Some people's probability table did not sum to 1 -- as probabilities always must (normalization)!
- Some people forgot the definition of conditional probability $p(H|O) = p(H,O)/p(O)$. Suggestion: this is just the standard definition $p(H,O) = p(H|O)p(O)$. Drill this into your minds, it is the nucleus of everything!
- Some people made little errors in Monty's probabilities.
- Some people find it counterintuitive that Monty opening a door can in any way make the probabilities change from equal (before

Announcements

- **mission training only** ("Homework 1") due before Friday's discussion section. See the link in CCLE Week 1.
- If any question in the exercises confuses you, mark it "Still confused, need help!" as your final status, and we'll try to provide additional materials on the question.

Three Mini-Languages for Inference

- **equations**: the chain rule and Bayes Law
- **pictures**: Venn diagrams
- **words**: a "Bayesian mini-language"

Bayes' Law In Words

Bayes Law is so central to inference that we will find ourselves referring to its various parts in nearly every sentence we utter when discussing an inference problem. So let's give each part a well-defined name, to eliminate any possible confusion about terms. Thus we can rewrite the symbolic form

$$p(H|O) = \frac{p(O|H)p(H)}{p(O)}$$

in words:

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Average Likelihood}}$$

A Recipe For Inference

- What is hidden (H)?
- What is observed (O)?
- What is the likelihood model p(O|H)?
- What is the prior p(H)?
- What is the set of all possible models H?
- What is the posterior p(H|O)?

What is hidden?

- Let's label the three doors *A*, *B* and *C*.
- Let's say you chose door *A*, and Monty opened door *B* to reveal that it was empty.
- Let's use the Greek letter δ for the hidden variable that indicates which door actually has the prize; i.e. $\delta = A$ means "door A has the prize". (In general in this text we will use Greek letters to designate hidden variables, on the principle that "it's Greek to me!").

Observables & Priors

- Let's use a minus sign to indicate that a door was observed to be empty; i.e. B^- means "door B was observed to be empty".
- This is just the unconditional probability for the prize to be hidden behind any of the three doors (with equal probability), i.e.

$$p(\delta = A) = p(\delta = B) = p(\delta = C) = 1/3$$

How would the prize's location affect the probability of our observation?

This is where we perceive the "connection" between O and δ. Let's consider the three possible doors.

- If door A has the prize, Monty can choose either B or C to reveal as empty, with equal probability. So $p(B^-|\delta = A) = 1/2$.
- If door B had the prize, of course there's no way it could be observed as empty, so $p(B^-|\delta = B) = 0$.
- Finally, if door C had the prize, Monty has no choice: he can't show A (you picked that one); he can't show C (it wouldn't be empty, and the game would be over), so $p(B^-|\delta = C) = 1$.

What is the total probability of our observation?

- Either you picked the right door ($\delta = A$) or you didn't ($\delta \neq A$).
- In the first case, the two remaining doors are equally likely to be chosen by Monty to be revealed to be empty.
- In the second case, both of the remaining doors are equally likely to have the prize, or conversely, equally likely to be opened by Monty and shown as empty.
- Either way, Monty is equally likely to open either of the two doors B or C.
- Thus, intuitively, $p(B^-) = 1/2$. (We will show how to calculate $p(O)$ rigorously later).

Bayes' Law Solution

So according to Bayes Law,

$$p(\delta = A|B^-) = \frac{p(B^-|\delta = A)p(\delta = A)}{p(B^-)} = \frac{\frac{1}{2}\frac{1}{3}}{\frac{1}{2}} = \frac{1}{3}$$

and

$$p(\delta = C|B^-) = \frac{p(B^-|\delta = C)p(\delta = C)}{p(B^-)} = \frac{1 \cdot \frac{1}{3}}{\frac{1}{2}} = \frac{2}{3}$$

So you are twice as likely to win by switching your choice to C, compared to staying with A! Monty's "observation" contains a lot of information about the prize location, and this is quantified directly by Bayes Law.

Normalization and Projection

- The probabilities of a set of disjoint subsets (i.e. all possible states of any random variable) must sum to 1.

$$\sum_X p(X) = 1$$

- This also defines a *projection*, i.e. eliminating a variable from a joint probability, by summing over all its possible values:

$$\sum_X p(X, Y) = \left(\sum_X p(X|Y) \right) p(Y) = p(Y)$$

Summation Equalities

Assume that a random variable B has disjoint possible values $\{b_1, \ldots, b_n\}$. Which of the following expressions if any are equal?

1. $\sum_{i=1}^{n} p(A|B = b_i)$.
2. $\sum_{i=1}^{n} p(A, B = b_i)$.
3. $\sum_{i=1}^{n} p(A|B = b_i) p(B = b_i)$.
4. $\sum_{i=1}^{n} \frac{p(A)}{p(B=b_i)}$.
5. $\sum_{i=1}^{n} p(A) p(B = b_i)$.
6. $p(A)$
7. $p(B)$

Summation Equalities Answer

2. $\sum_{i=1}^{n} p(A, B = b_i)$.
3. $\sum_{i=1}^{n} p(A|B = b_i)p(B = b_i)$.

5. $\sum_{i=1}^{n} p(A)p(B = b_i)$.
6. $p(A)$

Summation Equalities Errors

- Some people thought #1 sums to 1. No, normalization only applies to summing over the *subject* variable, not the *condition*. Beware of summing over the condition of a conditional probability!! This is a common error.
- Some people thought #1 sums to p(A). No, this sum isn't even a valid probability! E.g. consider *p(rains in Seattle|day of week)*.
- Many people realized 2 and 3 were equal, because $p(A, B) = p(A|B)p(B)$, but then didn't see the connection to 5, 6 ($p(A)$). This implies they didn't realize that summing over B eliminates it from the probability expression.

  This is a *really* important tool. It allows you to properly handle a hidden variable that links what you know (obs) and what you want to know (some other hidden variable). Just sum over all possible values of this "nuisance variable" and eliminate it from the equation!!
- Some people assumed A,B were independent, though the question made no such stipulation.

Statistical Independence

- *Statistical independence* is defined as a relationship between two random variables X, Y in which the conditional probability of one random variable is *always* equal to its unconditional probability:

$$p(X|Y) = p(X)$$

  (i.e. this holds true for all possible values of the random variables X, Y).
- This definition makes clear that knowing the value of Y tells you *nothing* about the likely value of X; its probabilities are totally unchanged.

Variable Independence

For a pair of random variables X, Y, you are told that $p(Y|X) = p(Y)$. Can you automatically conclude from this that $p(X|Y) = p(X)$?

Variable Independence Answer

Yes, this is always true. Recall that the chain rule is symmetric:

$$p(X, Y) = p(X|Y)p(Y)$$

$$p(X, Y) = p(Y|X)p(X)$$

So if $p(Y|X) = p(Y)$ then

$$p(X|Y)p(Y) = p(Y)p(X)$$

$$\text{So } p(X|Y) = p(X)$$

Variable Independence Errors

- Some people didn't consider the inherent symmetry of the intersection operation (i.e. the joint probability).

Independence Is Symmetric

- We can rewrite the statistical independence condition in *symmetric* form:

$$p(X, Y) = p(X)p(Y)$$

- Consider statistical independence in terms of its meaning for computational complexity:
  - say X, Y each have n states.
  - the joint probability table $p(X, Y)$ has $O(n^2)$ degrees of freedom.
  - there are $O(n^2)$ conditional probabilities $p(X|Y)$.
  - but the independent product $p(X)p(Y)$ has only $O(n)$ degrees of freedom -- a big reduction in computational complexity.

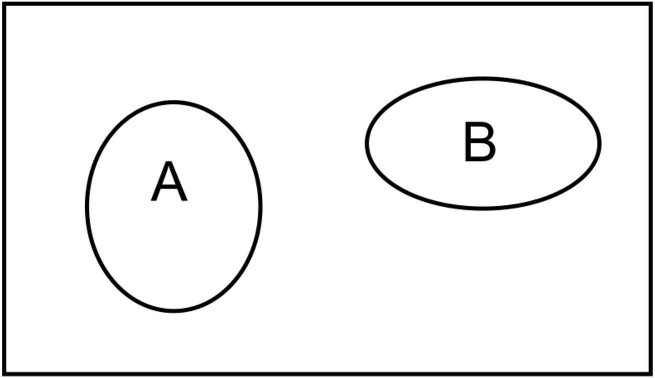
Random States vs. Random Variables

- A random variable is a specific way of slicing the set of all possibilities S into disjoint (non-overlapping) subsets, which are referred to as the *states* of the variable.
- Note that this definition makes explicit how a stochastic process emits *multiple* random variables: they are just *different* ways of dividing and labeling the same event set.
- A random variable is usually written as a capital letter e.g. X, whereas its possible *states* are usually written as lower case letters e.g. x. Thus $p(X = x)$ is the probability that random variable X will take the value x.

Drawing Multiple Variables At Once

- Note that multiple variables X, Y, Z simply mean *different* ways of subdividing S. Each time we sample S we obtain a value of *all* the variables $X, Y, Z, \ldots$ that we have defined on S. Depending on how we choose these subdivisions, there can be a lot, a little, or no correlation between two random variables. (Obviously, if we chose the same set of states for both variables, they would be perfectly correlated).
- Think of this as "throwing a dart at a dartboard" representing our total set. For each variable that we've defined on that set, that point lies in exactly one disjoint slice that corresponds to one value of that variable. So that one point gives us a value of *each* of the variables we've defined on that set.

Event Independence?



Are events A, B statistically independent? Justify your answer mathematically.

Event Independence? Answer

- First note that independence is only definable with respect to a pair of *variables*, not individual *events* or *states*. In other words, independence is a statement about the complete joint probability *function*. Simply saying that would be a completely valid answer.
- If we wanted to go deeper, we could answer the question by defining random variables based on A, B. Specifically, we define two binary variables $X \in \{A, \neg A\}$, $Y \in \{B, \neg B\}$. Now we can ask whether their joint probability mass function obeys independence, i.e.

  $$p(X, Y) = p(X)p(Y)$$

  The Venn diagram clearly fails this, e.g.

  $$p(X = A, Y = B) \neq p(X = A)p(Y = B)$$

Event Independence? Errors

- Some people asserted that $p(A|B) = p(A)$ in this Venn diagram, which implies a basic misunderstanding of what $p(A|B)$ *means*. Suggestion: review the definition of conditional probability until you can see clearly that this assertion does not fit this Venn diagram at all.
- Similarly, some people asserted $p(A \cap B) = p(A)p(B)$ here, implying they don't understand what a joint probability is (i.e. in this case that $p(A \cap B) = 0$).
- "They are independent since they have no overlap": mixing up the concepts of *independence* vs. *disjoint* (no intersection), which are totally different in meaning! Suggestion: Please go back and draw Venn diagrams of *independence* vs. *disjoint*.
- Similarly, some people seemed to define independence as meaning $p(A|B) = 0$. Suggestion: go back to the mathematical definition of independence.
- "yes, because an event in A has no effect on an event in B". Apparently not thinking about drawing a random event: if we draw

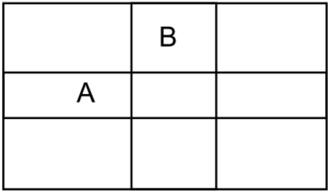
Event Independence? Errors

- "yes, because an event in A has no effect on an event in B". Apparently not thinking about drawing a random event: if we draw a point in A, we know it can't be in B, and vice versa, so actually they are perfectly *anti-correlated*, not independent.
- mixing up the concepts of *states* vs. *variables*. Independence is by definition a statement about *variables*, not states. Watch out for mixing these up! A state is just any subset of the probability space; A variable is "the whole picture" of a way of slicing up the entire space. Suggestion: draw a Venn diagram representing a single state, vs. a Venn diagram representing a variable.
- Some people said "we do not have enough info because we don't know the correlation between A and B". *Au contraire*, the Venn diagram shows they are perfectly anti-correlated. Suggestion: review the concepts of correlation and Venn diagrams until this is clear.

Venn Diagram of Independence

Draw a Venn diagram representing two *independent* variables. To keep the diagram clear and simple, restrict each variable to just three states, and draw the diagram so *area* represents (i.e. is proportional to) probability.

Venn Diagram of Independence Answer



The X variable is shown as being split into three discrete values on the x-axis. The Y variable is shown as being split into three discrete values on the y-axis. The probability of a given X value or (X,Y) pair is shown by the area of that region in the figure. For example, the cell labeled A represents the joint probability of the intersection $(X = x_1, Y = y_2)$, and B represents $(X = x_2, Y = y_3)$. Since the probability (area) of any given (X,Y) pair is simply proportional to the product of the probabilities (area fractions) of that value of X and of that value of Y, they are independent.

Venn Diagram of Independence Errors

- Some people said the space should be split in two disjoint regions, one representing X (split into its three values), and the other representing Y (also split into three values). NO, each random variable covers the entire space (i.e. its three states sum to the entire space).
- "One variable should be a subset of the space covered by the other variable". NO, each random variable covers the entire space.
- Some people simply weren't sure how to translate $p(X, Y) = p(X)p(Y)$ into a graphical representation, implying that they weren't thinking of this product relation in terms of *orthogonality*.