INFO 2301: Quantitative Reasoning for Information Science

Abe Handler
Department of Information Science
University of Colorado, Boulder

November 17, 2021

Contents

1 Sets

A **set** is an unordered collection of items, without duplicates.

People follow a number of conventions when talking about sets.

- The items in a set are often called “elements”
- We use capital letters to name sets
- We show the elements in a set using curly brackets
- A set can include any kind of element (e.g. strings, or kinds of apple)

Example 1.1. We can write the integers between 0 and 3 (including 0 and 3) as $A = \{0, 1, 2, 3\}$

Example 1.2. Because sets do not have order, if $B = \{0, 1\}$ and $C = \{1, 0\}$, then $C = B$.

Example 1.3. A set of strings, $P = \{$“Denver”,“Boulder”,“Broomfield”$\}$

- If a is an element in X then, we use the notation $a \in X$
- If $a \in X$ then we say that “a is in X”
- We write this: `$a \in X$`

- If $a \notin X$ then we say that "a is not in X"

- We write this: `$a \notin X$`

Example 1.4. Let $C = \{1, 2\}$. Then $1 \in C$ and $6 \notin C$

1.1 Subsets

If all elements of A are elements of B then we say that A is a subset of B. We use the notation $A \subset B$ (written `$ A  \subset B $`) to indicate that A is a subset of B.

Example 1.5. Let $A = \{2, 1\}$ and $B = \{1, 2, 5\}$. Then $A \subset B$

In this class, when we use $A \subset B$, we will usually assume that there is at least one element of B that is not an element of A. For instance, in Example 1.5, $A \subset B$ because $5 \in B$ and $5 \notin A$. If we wanted to be a little more precise, we could say that A is a "proper subset" of B, which means that we know A is not equal to B (because there is an element in B that is not in A).

You will also sometimes see $A \subseteq B$, which means that at least one of the following things is true: either A is a proper subset of B, or A is equal to B. The $\subseteq$ symbol is written `$a \subseteq X$`.

Example 1.6. Let $A = \{1, 2, 3\}$ and $B = \{1, 2, 3\}$. Then $A \subseteq B$ because every element of A is in B so $A = B$. However, A is not a proper subset of B, because there is no element of B that is not in A.

Example 1.7. Let $A = \{1, 3\}$ and $B = \{1, 2, 3\}$. Then $A \subseteq B$ because one of the following is true: either A is proper subset of B or A is equal to B. (It's the former.) Additionally, $A \subset B$ because A is a subset of B.

Note: you can use the symbol `$\subsetneq$` to show that A is not a subset of B, i.e. $A \subsetneq B$.

Example 1.8. Let $A = \{\mathit{dogbreeds}\}$ and $B = \{\mathit{catbreeds}\}$. So $A \subsetneq B$.

1.2 Intersections, unions and complements

- We use the notation $A \cap B$ (written `$ A  \cap B $`) to indicate all of the elements that are in *both* A and B. This is called the *intersection* of A and B

- We use the notation $A \cup B$ (written `$ A  \cup B $`) to indicate all of the elements that are in A or B. This is called the *union* of A and B.

- To understand intersections and unions it is often helpful to draw venn diagrams.

Example 1.9. Let $A = \{1, 2, 3\}$ and $B = \{2, 3, 7\}$. Then $A \cap B = \{2, 3\}$ and $A \cup B = \{1, 2, 3, 7\}$

- The complement of a set A is the set of all elements that are not in A
- The complement of a set requires defining U, the universal set of all possible elements
- We write the complement of A as $\overline{A}$ (written `$ \overline{A} $`
- We can also write the complement as A^c (written `$ A^c $`)

Example 1.10. Let $U = \{1, 2, 3\}$ and $A = \{2\}$. Then $\overline{A} = \{1, 3\}$

1.3 Cardinality

The number of items in a set is called the *cardinality* of the set. We use the notation $|F|$ (written `$ \vert F \vert $`) to indicate the cardinality of a set.

Example 1.11. Let $F = \{1, -2, 42\}$, then $|F| = 3$

Example 1.12. Let $B = \{7, 42\}$, then $|B| = 2$

Example 1.13. Let $A = \{1, -2, 42\}$ and $B = \{1, 4\}$ then $|A \cup B| = |\{1, -2, 42, 4\}|$ $= 4$

2 Functions

- A function maps elements from one set to another.
- The first set is called the domain.
- The second set is called the range.
- A function maps each element from the domain to exactly one element in the range.
- We write this $f : A \rightarrow B$ where f is a function mapping elements of A to B (written `$f: A \rightarrow B $`)

Example 2.1. Let A be a set $\{0, 1, 2, 3\}$ and let f be a function mapping every element of A to -19. The domain of f is A and the range is $\{-19\}$

Example 2.2. Let A be the set of all integers and let B be the set of all integers. The `addOne` function maps each element in A to the element in B that is exactly one greater. For instance, the `addOne` function maps 4 to 5.

3 Vectors

- A vector is a list of numbers.
- You can think of a vector as generalizing a single number in one dimension (called a scalar) to a number in multiple dimensions
- We denote vectors with bold, lower-case letters and triangle brackets, such as $\mathbf{x} =< 1, 2 >$ (written `$  \mathbf{x}=<1,2> $`)
- Each number in the vector is called a "component" and represents a dimension of the vector
- We can use subscripts to refer to components of a vector.

Example 3.1. If $\mathbf{x} =< 6, 9 >$ then $\mathbf{x}_1 = 6$ (assuming indexing starts at one)

Sometimes, when taking about vectors, people describe the ordinary numbers you are familiar with as "scalars." For instance, 7 is a scalar. And so is 14. You can also think of scalars as 1-D vectors, or think of vectors as generalizations of scalars in multiple dimensions.

Example 3.2. $\mathbf{x} =< 7 >$ is a vector in a 1-dimensional space. If you represent that space as a number line, then the vector points from zero to seven along the single, 1-D number line. In this sense $\mathbf{x}$ is very, very similar to the ordinary scalar 7.

Example 3.3. $\mathbf{x} =< 7, 2 >$ is a vector in a 2-dimensional space. The first component specifies how far the vector goes in the first dimension, i.e. 7 units. So in this sense $\mathbf{x} =< 7, 2 >$ extends the ordinary scalar 7 by adding information about an additional second dimension.

People can only really think in 2D or 3D but vectors can be in any number of dimensions. Another way express that is to say that vectors generalize to any arbitrary number of dimensions. This is very handy for data science. People can't really think in, say, 5 dimensions. But everything that is true about vectors in 2 dimensions is also true about vectors in 5 dimensions. Thus we can use vector math to reason about and draw conclusions from high-dimensional data, such as if a person's medical results on 5 different lab tests. In data science, it is common to have vectors with hundreds or even thousands of components.

Example 3.4. $\mathbf{x} =< 1, 22, -2, -3, 0 >$ is a vector in a 5-dimensional space, representing a person's results on 5 different medical tests. For instance $\mathbf{x}_1$ might represent their result on a binary diagnostic test, which returns 1 if the person has a disease, and 0 otherwise.

3.1 Vector addition

Vector addition is an operation which takes two input vectors and returns one output vector. If $\mathbf{x} =< x_1, x_2 ... x_n >$ and $\mathbf{y} =< y_1, y_2 ... y_n >$ then $\mathbf{x} + \mathbf{y} =< x_1 + y_1, x_2 + y_2 ... x_n + y_n >$.

Example 3.5. $\mathbf{x} =< 3, 4 >$ and $\mathbf{y} =< -1, 2 >$ then $\mathbf{x} + \mathbf{y} =< 2, 6 >$

You can think of vector addition as generalizing scalar addition. Scalar addition takes two floats a and b and returns another float, that equal to moving a and b along a single-dimensional number line. Vector

3.2 Summation notation

Summation notation is important for vector operations, and in many other areas of math.

- To indicate the sum of a sequence of items, we write $\Sigma_{i=0}^{N} x_i$ where x_i is the ith item in a sequence.
- You can read this as summing over a sequence of items, from item 0 to item N.
- You can think of i as an index into the sequence, in the same way you may think of i as an index into an array.
- Usually, in 2301, the sequence being indexed is positive integers: 1,2,3,4 ... etc. So x_1 is usually 1, x_2 is usually 2, etc. You can assume we are indexing into the positive integers unless otherwise specified.
- We write sigma notation in LaTeX as `$ \Sigma_{i=0}^{N}x_i $`

Example 3.6. You can also use summation notation without indexing into a sequence, e.g. $\Sigma_{n=1}^{N=3}(n + 1) = (1 + 1) + (2 + 1) + (3 + 1) = 2 + 3 + 4$.

You can also use this notation to describe the sum of items in a sequence, after applying some operation.

Example 3.7. If $x = [1, 2, 3, 4, 5]$ then $\Sigma_{i=1}^{5} x_i^2 = 1^2 + 2^2 + 3^2 + 4^2 + 5^2 = 1 + 4 + 9 + 16 + 25 = 55$

Example 3.8. Note that you should read the i as indexing into an array. So if you have an array $x = [2, 3, 4]$ then $\Sigma_{i=1}^{3} x_i = x_1 + x_2 + x_3 = 2 + 3 + 4$

3.3 Euclidean norm

The Euclidean norm of a vector is $||x||_2 = \sqrt{\Sigma_{i=1}^{N} x_i^2}$. You can think of the Euclidean norm as the size of a vector.

Example 3.9. If $\mathbf{x} =< -2, 2 >$ then $||x||_2 = \sqrt{-2^2 + 2^2} = \sqrt{4 + 4} = \sqrt{8}$

Note: There are also other norms, which offer other ways of defining the size of a vector.

3.4 Normalizing a vector

- To normalize a vector, divide each component by the (Euclidean) norm. This creates a new “unit vector” with a norm of 1.
- People sometimes write unit vectors with “hats” over them, e.g. $\hat{\boldsymbol{b}}$ (written `$ \hat{\boldsymbol{b}} $`, where `$ \boldsymbol $` means to write the symbol in bold).

Example 3.10. If $\mathbf{x} =< -2, 2 >$ then $||x||_2 = \sqrt{-2^2 + 2^2} = \sqrt{4 + 4} = \sqrt{8}$. The unit vector of $\mathbf{x}$ is $\hat{\mathbf{x}} =< \frac{-2}{\sqrt{8}}, \frac{2}{\sqrt{8}} >$. The Euclidean norm of the unit vector is $\sqrt{(\frac{-2}{\sqrt{8}})^2 + (\frac{2}{\sqrt{8}})^2} = \sqrt{\frac{4}{8} + \frac{4}{8}} = 1$, so you know you have a unit vector.

3.4.1 You can normalize any vector

Any vector can be converted into a unit vector. Here is an argument for this claim in 2D. It is easy to extend to any dimension. It is OK to skip this part if you want to just trust that you can always normalize. Assume we have a Euclidean norm of N. The following is true by definition in 2D.

$$a^2 + b^2 = N^2 \tag{1}$$

Divide each side by N^2

$$\frac{a^2}{N^2} + \frac{b^2}{N^2} = 1 \tag{2}$$

Take the root of each side

$$\sqrt{\frac{a^2}{N^2} + \frac{b^2}{N^2}} = \sqrt{1} \tag{3}$$

Simplify, recalling that the square root of 1 is 1.

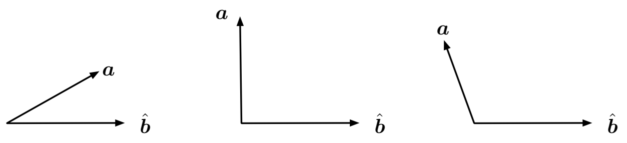
$$\sqrt{(\frac{a}{N})^2 + (\frac{b}{N})^2} = 1 \tag{4}$$

Therefore, by definition, if you divide each component by N you get a normalized vector.

3.5 Dot product

There are two ways to think about the dot product. You can think about it geometrically (in pictures) or algebraically (in symbols). It’s better to build geometric intuition than to plug and chug through the algebraic version, but both are important to know really well.

- Algebraically, the dot product is defined as $\boldsymbol{a} \cdot \boldsymbol{b} = \sum_i \boldsymbol{a}_i \boldsymbol{b}_i$.
- Geometrically, the dot product is defined as $\boldsymbol{a} \cdot \boldsymbol{b} = ||\boldsymbol{a}||\ ||\boldsymbol{b}|| \cos\theta$.



(a) a (b) b (c) b

Figure 1: The dot product is positive when the angle θ between $\boldsymbol{a}$ and $\hat{\boldsymbol{b}}$ is less than 90 degrees (left). It is zero when $\theta = 90°$. It is negative when θ is greater than $90°$. Try plugging in pairs like this to the geometric definition of the dot product.

Example 3.11. If $\mathbf{x} =< -2, 2 >$ and then $\mathbf{y} =< -1, 7 >$ then applying the algebraic definition we have $\boldsymbol{x} \cdot \boldsymbol{y} = \sum_i \boldsymbol{x}_i \boldsymbol{y}_i = -2 * -1 + 2 * 7 = 2 + 14 = 16$

Example 3.12. If $\mathbf{x} =< 1, 0 >$ and then $\mathbf{y} =< 0, 1 >$ then by the algebraic definition we have $\boldsymbol{x} \cdot \boldsymbol{y} = \sum_i \boldsymbol{x}_i \boldsymbol{y}_i = 1 * 0 + 0 * 1 = 0$. But also in this case the angle between $\mathbf{x}$ and $\mathbf{y}$ is 90 degrees (draw it!). Because the cosine of 90 degrees is zero, $\boldsymbol{a} \cdot \boldsymbol{b} = ||\boldsymbol{a}|| \; ||\boldsymbol{b}|| \cos{(90)} = ||\boldsymbol{a}|| \; ||\boldsymbol{b}|| \; * 0 = 0$. Hence the two definitions give the same answer (which they should!).

Loosely, the dot product of $\boldsymbol{a}$ and $\boldsymbol{b}$ measures the extent to which $\boldsymbol{a}$ goes in the direction of $\boldsymbol{b}$. More precisely, the dot product of $\boldsymbol{a}$ and $\hat{\boldsymbol{b}}$ (i.e. the unit vector of $\boldsymbol{b}$) is defined as the projection of $\boldsymbol{a}$ onto $\hat{\boldsymbol{b}}$. You can think of the projection as the shadow $\boldsymbol{a}$ casts onto $\hat{\boldsymbol{b}}$, which is the adjacent size of a right triangle. See Figure 2.

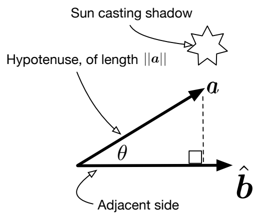


Figure 2: The dot product of $\boldsymbol{a}$ and $\hat{\boldsymbol{b}}$ is the projection of $\boldsymbol{a}$ onto $\hat{\boldsymbol{b}}$. You can think of this as $\boldsymbol{a}$ casting a shadow onto $\boldsymbol{b}$ to form a right triangle with hypotenuse of length $||\boldsymbol{a}||$.

Recall from trigonometry that $cos(\theta) = \frac{\text{adjacent}}{\text{hypotenuse}}$. In this case, the hypotenuse is $||\boldsymbol{a}||$ so we have $cos(\theta) = \frac{\text{adjacent}}{||\boldsymbol{a}||}$ so we can rearrange to get $cos(\theta)||\boldsymbol{a}|| = \text{adjacent}$. So by definition $\boldsymbol{a} \cdot \hat{\boldsymbol{b}} = \boldsymbol{a} \cdot \frac{\boldsymbol{b}}{||\boldsymbol{b}||} = cos(\theta)||\boldsymbol{a}||$. If we rearrange we get the geometric definition of the dot product $\boldsymbol{a} \cdot \boldsymbol{b} = cos(\theta) * ||\boldsymbol{a}|| * ||\boldsymbol{b}||$

Note that the geometric definition implies that when the angle between the vectors is acute, the dot product is positive. When it is obtuse the dot product is negative. To see this, consider the value of $cos(\theta)$ in these cases. Note that when two vectors are at a right angle, the dot product is zero; the vectors are said to be orthogonal.

4 Booleans

A Boolean variable can take on the value True or False. A Boolean expression is a collection of Boolean variables, joined by logical operators.

4.1 NOT

The simplest Boolean operator is NOT. The NOT operator takes a single Boolean operand. It flips the value of the Boolean variable (so True becomes False, and vice versa.) For instance, A might be Boolean variable equal to True. Then NOT A is False. We also write this as $\neg A$ (`$ \neg $`).

4.2 AND

The Boolean operator AND takes two Boolean variables as operands. The operator returns True if both variables are True. We write this $A \wedge B$ (`$ A \wedge B $`)

4.3 OR

The Boolean operator OR takes two Boolean variables as operands. The operator returns True if either variables is True. We write this $A \vee B$ (`$ A \vee B $`)

Example 4.1. If A=True and B=True then $A \wedge B$ is True

Example 4.2. If A=True and B=False then $A \wedge B$ is False

Example 4.3. If A=True and B=False then $A \vee B$ is False

Example 4.4. If A=True and B=False then $A \wedge \neg B$ is True because $\neg B$ is True and the AND of two true variables is True

4.4 Evaluation

Evaluating a Boolean expression means determining if the expression is True or False, based on the values of its variables. To evaluate a Boolean expression you should fill in the value of each variable and simplify.

Example 4.5. If A=True and B=False and C=True then $(A \vee B) \wedge C = (T \vee F) \wedge T = T \wedge T = T$

5 Probability

5.1 The sample space

The sample space Ω (written `$\Omega$`) is the set of all outcomes of an experiment. An event A is a subset of the sample space.

Example 5.1. If we toss a coin twice then Ω=$\{HH, HT, TH, TT\}$

Example 5.2. If we toss a coin twice then Ω=$\{HH, HT, TH, TT\}$. The event "at least one tails" is $A = \{HT, TH, TT\}$. Notice that $A \subset \Omega$.

5.2 Probability distribution

A probability distribution is a function that maps events to real numbers between 0 and 1, and that satisfies the following three properties

- $p(A) \geq 0$ for all A
- $p(\Omega) = 1$
- If A and A' are disjoint (i.e. their intersection is $\emptyset$) then $p(A \cup A') = p(A) + p(A')$

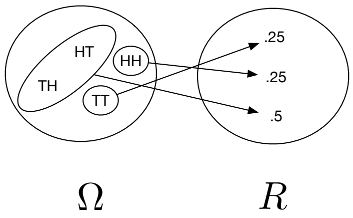


Figure 3: An experiment flipping a fair coin twice. This probability distribution maps three events in Ω to a number between 0 and 1.

There are two interpretations of a probability distribution

- In the long run, if you keep running an experiment, the probability of a given event A is the fraction of times you observe A, among all times you run the experiment. For instance, if you keep flipping a fair coin forever, you should expect that, in the long run, half of the time you will get heads. This is sometimes called the **frequentist** interpretation.
- Another interpretation of a probability distribution is that the distribution reflects your subjective belief about what will happen if you run an experiment in the future. For instance, if you believe (based on observing data) that you will get heads about half of the time the fact that $p(H) = .5$ (i.e. the probability of heads is 50%) reflects your belief that about what will happen if you flip a coin. This perspective is sometimes called the **Bayesian** interpretation.

5.3 The uniform distribution

If all events are equally probable, we say that the distribution is uniform. If a distribution is uniform, then probability of any event $A \subset \Omega$ is $p(A) = \frac{1}{|\Omega|}$.

Example 5.3. If you roll a fair 6-sided die, all outcomes are equally likely so the distribution over outcomes is uniform. The probability of the event 1 is $p(\{1\}) = \frac{1}{|\Omega|} = \frac{1}{6}$.

5.4 Determining a distribution from data (aka "learning")

In the previous examples, we have largely known how the data was generated. For instance, we know that a fair die will land on the number 3 roughly $\frac{1}{6}$ of the time (if we keep rolling the die forever). But usually in nature you *don't* know how the data was generated. You just see data and you have to make an educated guess about how it was created.

If you have heard people mentioning **machine learning**, part of what they are talking about is a set of computational methods for reasoning about an (unobserved) probability distribution from observed data. Outside of computer science, you will also hear the process of reasoning about an unobserved distribution from data described as **statistical inference**.

Example 5.4. For example, say you observe 10 flips of a coin and get TTTTTHTTTTT. Based on this data, do you think the probability distribution (i.e. the long run probabilities of getting heads or tails) is uniform? Well, there are two outcomes $\Omega = \{\text{H, T}\}$. If a distribution is uniform then we should expect that $p(\text{H}) = \frac{1}{|\Omega|} = \frac{1}{2}$ and $p(\text{T}) = \frac{1}{|\Omega|} = \frac{1}{2}$.

In this example, in the observed data there are 10 flips and only 1 heads. It is possible to get 9 out of 10 tails from a fair coin, but it is really unlikely. So you might make an educated guess that $p(\text{H}) \neq \frac{1}{2}$ and that the distribution is not uniform. If you were to keep flipping the coin 10,000 times and see that only 1 out of 10 flips is heads, you can be pretty confident (but not totally certain) that the distribution is not uniform.

You can think of generating data $\mathcal{D}$ based on some (unseen) θ and then trying to infer an estimate of θ (denoted) $\hat{\theta}$ based on $\mathcal{D}$.

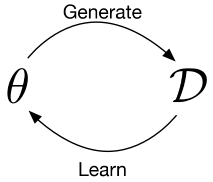


Figure 4: We imagine that data $\mathcal{D}$ is generated from an unseen parameter θ. Based on observed data $\mathcal{D}$ we try to infer or learn the value of the unseen θ.

5.5 Independence

Two events A and B are independent if $P(A \cap B) = P(A)P(B)$. Intuitively, if the outcome of A does not affect the outcome of B then A and B are independent.

Example 5.5. If you flip a coin twice, then the value of the first flip does not effect the value of the second flip. Thus the flips are independent.

Example 5.6. On the other hand, your height and weight are likely not independent; if you are a taller person you likely have a larger weight than a shorter person. We say that these events are dependent (i.e. not independent).

Example 5.7. Let's say we toss a fair coin five times. What is the probability of getting at least 1 head? Well, because probability sums to 1, the probability of at least one head is 1 minus the probability of all tails. If the fail flips are independent, then this is 1 minus the probability of tails on the first flip times the probability of tails on the second flip times the probability of tails on the third flip.

5.6 Conditional probability

- The conditional probability of event A is the probability of observing event A given that you observed event B.
- We write conditional probability using the notation $p(A|B)$ (i.e. `$ p(A \vert B) $`). You should read this as the probability of A given B.
- Formally, $p(A|B) = \frac{p(A \cap B)}{p(B)}$

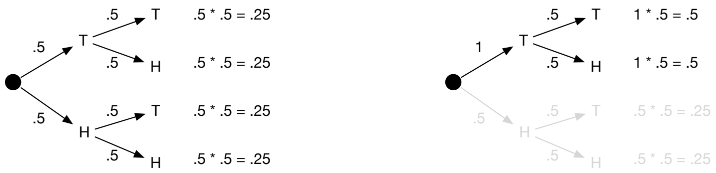
Example 5.8. For instance, in general, taller people tend to weigh more than shorter people. So if you know that someone weighs over 200 pounds, it's more likely that they will be over 6 feet tall (compared to a person who weighs less than 200 pounds). Thus, the probability of observing event A (i.e. a person who is over 6 feet tall) is higher if you first observe event B (i.e. observing a person who weighs more than 200 pounds).

5.6.1 Independence revisited, after conditional probability

Intuitively, if two events A and B are independent, then the value of A does not affect the value of B. We can reconsider independence in light of conditional probability. $p(B|A)$ gives the probability of B given A. If A and B are independent then we should not expect the value of B to depend on A. We can show this in the math.

By definition:

$$p(B|A) = \frac{p(B \cap A)}{p(A)} \tag{5}$$



(a) The probability of observing two tails on two flips is .25

(b) The probability of observing two tails, given that the first flip is T is .5

Figure 5: Probability of getting two tails on two flips of a fair coin. Note that the events are independent so $p(\mathrm{T} \cap \mathrm{T}) = p(\mathrm{T})\mathrm{p}(\mathrm{T})$ and p(T | T) =p(T)

If B and A are independent, then $p(B \cap A) = p(B)p(A)$. Thus we can rewrite the above:

$$p(B|A) = \frac{p(B)p(A)}{p(A)} \tag{6}$$

The $p(A)$ terms cancel (these just represent numbers after all) so if A and B are independent we have:

$$p(B|A) = p(B) \tag{7}$$

Note that this is NOT true if A and B are dependent (i.e. not independent).

5.7 Multiplication rule for dependent events

Whenever we have two events A and B we multiply their probabilities to get the probability that A occurs and B occurs, written $p(A \cap B)$.

Assuming that A and B are dependent, then what is $p(A \cap B)$? One way to think about this is that first event A happens then event B happens and we need to multiply the probabilities of the events together to get the probability of both events, $p(A \cap B)$. The probability of event A can be written $p(A)$. The probability of event B given that A has already occurred can be written $p(B|A)$. Thus $p(A \cap B) = p(A)p(B|A)$. This is sometimes called the multiplication rule for dependent events. What do you think the multiplication rule for independent events might be?

5.8 Law of total probability

If you have sets B_1, B_2 ... B_N that are all disjoint (i.e. don't intersect) then the law of total probability states:

$$p(A) = \sum_n p(A|B_n)p(B_n)$$

The law of total probability is maybe best explained with a picture, as in Figure 6.

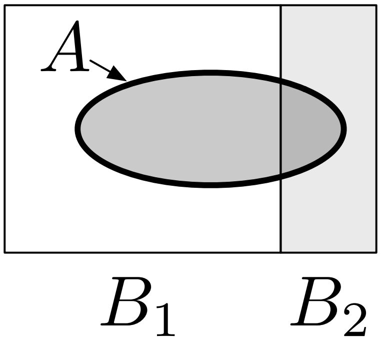


Figure 6: The law of total probability: $p(A) = p(A|B_1)p(B_1) + p(A|B_2)p(B_2)$. The probability of the oval A is equal to the probability of region B_1 times the probability of A given that you are in the region B_1, plus the probability of region B_2 times the probability of A given that you are in the region B_2.

Example 5.9. Sixty percent of students in a district go to high school A and forty percent of students in a district go to high school B. Ninety percent of students from high school A graduate in 4 years and eighty percent of students form high school B graduate in 4 years. What is the overall probability that a student from the district? From the law of total probability, the probability is $.6 * .9 + .4 * .8$.

Example 5.10. An example from natural language processing. Say ten percent of stories on ESPN.com are about baseball and ninety percent of stories on ESPN.com are about football. One out of two stories about football includes the word "touchdown", and one out of one hundred stories about baseball includes the word "touchdown". You can apply the law of total probability to compute the overall probability of the word "touchdown". In NLP, this is called a *mixture model* of text.

5.9 Random variables

A random variable $X : \Omega \to \mathbb{R}$ is a function that maps each element in the sample space to a real number. Sometimes we abbreviate random variables as "r.v."

Example 5.11. If you flip a coin twice, then $\Omega = \{HH, HT, TH, TT\}$. Let X be a r.v. equal to the number of heads in a sequence. Thus X maps HH to the number 2. X maps TT to the number 0.

5.10 Expected value

We write the expected value as $E[X]$. Intuitively, you can think of the expected value as what you should expect the value of X to be. Formally, for a discrete r.v. (we will only focus on discrete variables in this class) the definition is:

$$E[X] = \Sigma^X p(x_i) * x_i \tag{8}$$

where $p(x_i)$ is the probability of some outcome and x_i is the value of that outcome.

Example 5.12. A friend offers to bet on the Broncos game on Sunday. If the Broncos win, you get $100. If the Broncos lose, you pay $50. Say that based on the last three years of Broncos games there is a 25% chance the Broncos win and a 75% chance the Broncos lose. The expected value of the proposed bet is .25 * -50 + .75 * 100 = 62.5. You should take the bet because you should expect to win $62.5.

5.10.1 A connection between average and expected value

What you are used to calling the "average" can be thought of a kind of (approximated) expected value. This may be shown most clearly with an example. Say we roll a six-sided die N times. The average would be $\frac{1}{N}\Sigma x_i$ where x_i is the value of a roll of the die (e.g. 6 or 4). If you get a 1 N_1 times a 2 N_2 times etc., then the average can be written

$$Avg = \frac{1}{N}\Sigma x_i = \frac{1}{N}\big[N_1 * 1 + N_2 * 2 + N_3 * 3 + N_4 * 4 + N_5 * 5 + N_6 * 6\big]$$

where x_i is a value of X. If you rearrange the sum by "pushing in" the N then we get:

$$Avg = \frac{1}{N}\Sigma x_i = \frac{N_1}{N} * 1 + \frac{N_2}{N} * 2 + \frac{N_3}{N} * 3 + \frac{N_4}{N} * 4 + \frac{N_5}{N} * 5 + \frac{N_6}{N} * 6$$

where for instance $\frac{N_3}{N}$ is the fraction of times we roll a 3.

By definition (according to the frequentist view), $p(3)$ is the number of times you get a 3 as you run N experiments, as N gets very large. Hence $\frac{N_3}{N}$ is an approximation of $p(3)$. When we estimate something in statistics we add a "hat" over it. Hence we can write $\frac{N_3}{N} = \hat{p}(3)$. If we replace $\frac{N_3}{N}$ as $\hat{p}(3)$ etc. in the equation we get

$$Avg = \frac{1}{N}\Sigma x_i = \hat{p}(1) * 1 + \hat{p}(2) * 2 + \hat{p}(3) * 3 + \hat{p}(4) * 4 + \hat{p}(5) * 5 + \hat{p}(6) * 6$$

which is an approximation of $E[X] = \Sigma p(x_i)x_i$.

5.11 Variance

Some random variables take values that are spread out. Other random variables take values in a narrower range.

Example 5.13. House cats usually way between 5 and 20 pounds. Pet dogs usually weigh between 10 and 100 pounds. Say you have a sample of pet cats and dogs. If X is a r.v. that maps a dog in a sample to its weight, and Y is an r.v. that maps a cat in a sample to its weight, then X is more spread out than Y.

- The variance offers a way to quantify how spread out a variable is.
- A variance is an expected value.
- Specifically, it is the expected value of how far a variable is from the mean.
- Formally, the variance of a discrete r.v. is $Var[X] = E[(X - E[X])^2] = \sum p(x_i)(X - E[x])^2$
- The standard deviation σ is the square root of the variance

  i.e. $\sqrt{Var(X)} = \sigma$

Example 5.14. You flip a coin twice. Let X be a r.v. mapping to the number of heads across two tosses. The sample space is $\Omega = \{HH, HT, TH, TT\}$ and $p(X) = \frac{1}{4}$. Recall that the r.v. X is a function. So, for instance, $X(HH) = 2$ and $X(HT) = 1$.

- Let's start by computing the expected value, which we need to calculate the variance. By definition, we get the following:

  $E[X] = \sum p(HH)X(HH) + p(HT)X(HT) + p(TH)X(TH) + p(TT)X(TT)$
- If we plug in numbers for probabilities we get the following:

  $E[X] = \sum \frac{1}{4}X(HH) + \frac{1}{4}X(HT) + \frac{1}{4}X(TH) + \frac{1}{4}X(TT)$
- If we plug in numbers for values of X we get the following:

  $E[X] = \sum \frac{1}{4} * 2 + \frac{1}{4} * 1 + \frac{1}{4} * 1 + \frac{1}{4} * 0$
- If we simplify we get the following:

  $E[X] = \sum \frac{2}{4} + \frac{1}{4} + \frac{1}{4} + 0 = 1$
- Thus $E[X] = 1$, meaning that if we flip a coin twice, on average, we will get 1 head.
- Now that we know $E[X]$ we can ask: what is the variance? We start by applying the definition.
- $Var[X] = \sum p(x_i)(X - E[x])^2$

- Thus...
- $Var[X] = \sum p(HH)(X(HH)-1)^2 + p(HT)(X(HT)-1)^2 + p(TH)(X(TH)-1)^2 + p(TT)(X(TT)-1)^2$
- If we plug in numbers for probabilities we get the following:

  $Var[X] = \sum \frac{1}{4}(X(HH)-1)^2 + \frac{1}{4}(X(HT)-1)^2 + \frac{1}{4}(X(TH)-1)^2 + \frac{1}{4}(X(TT)-1)^2$
- If we plug in values for random variables, we get the following:

  $Var[X] = \sum \frac{1}{4}(2-1)^2 + \frac{1}{4}(1-1)^2 + \frac{1}{4}(1-1)^2 + \frac{1}{4}(0-1)^2$
- If we simplify we get the following:

  $Var[X] = \sum \frac{1}{4}(1)^2 + \frac{1}{4}(0)^2 + \frac{1}{4}(0)^2 + \frac{1}{4}(-1)^2$
  $Var[X] = \sum \frac{1}{4} * 1 + \frac{1}{4} * 1$
  $Var[X] = \frac{1}{2}$
- The expected squared deviation from the mean (i.e. the variance) is $\frac{1}{2}$. This is a quantitative measure of how spread out X is.

5.12 Bayes rule

Bayes rule tells you the probabilty of some event B, given that you know that some other event A has happened. Usually when you apply Bayes rule you know how to compute $p(A|B)$ but not $p(B|A)$. Thus you apply the law of conditional probability like this:

$$\begin{aligned} p(B|A) &= \frac{p(B \cap A)}{p(A)} && \text{law of conditional probability} \\ &= \frac{p(A|B)p(B)}{p(A)} && \text{via multiplication law} \end{aligned}$$

this gives Bayes rule $p(B|A) = \frac{p(A|B)p(B)}{p(A)}$. Note that $p(A) = \sum p(A|B')p(B')$, by the law of total probability. To remember Bayes rule, it helps to think of "flipping the conditional" from $p(A|B)$ to $p(B|A)$.

Note: In Bayesian statistics, the terms on the right have special names. $p(A)$ is called the *evidence* (sometimes also the "normalizer"), $p(A|B)$ is called the *likelihood* and $p(B)$ is called the *prior*. The overall probability is called the *posterior*.

6 Permutations and combinations

6.1 Permutations

Permutations enumerate the items in a set in a particular order. Recall that sets do not have order.

Example 6.1. Let A be the set $\{45, 67, 99\}$. The set does not have an order. $67, 45, 99$ is a permutation of the set, which has an order.

If there are n items in a set and we permute k of the items, there are $\frac{n!}{(n-k)!}$ possible permutations, where the explanation point denotes the factorial function. You can think of this as follows: there are n options for the first item in the permutation, $n-1$ options for the second item in the permutation. In total there are $n*(n-1)*(n-2)...(n-k+1)$ permutations. In the formula $\frac{n!}{(n-k)!}$ the denominator cancels out the last $(n-k)$ terms of the $n!$ function.

Example 6.2. How many ways are there to pick a first place winner and a second place from 10 contestants? There are 10 people who can come in first place and for each of the 10 people, there are 9 who can come in second place. Thus there are $10*9 = 90$ permutations. We can also compute this via $\frac{n!}{(n-k)!}$ $= \frac{10!}{8!} = 90$ where n is 10 and k is 2.

We sometimes write the number of permutations as $P(n,k) = \frac{n!}{(n-k)!}$ which reads: how many ways are there to choose k items from a set of size n, where order matters.

6.2 Combinations

Combinations enumerate the items in a set where order does not matter.

Example 6.3. For instance, if you play the traditional lottery and your numbers come up, it does not matter which order you picked the numbers. You still win the lottery, regardless of the order in which you pick the numbers.

One way to think about counting combinations is to start by thinking of permutations, and then consider how to remove duplicates. For instance, if you have a permutation of k items, there are $k!$ possible permutations of the k items. So you need to divide the total number of permutations by $k!$.

Example 6.4. How many ways are there to pick two winners from 10 contestants? Well you can pick one of the 10 people first, then pick one of nine people second. However, in this case, the order of the winners does not matter. Thus picking person A and person B is the same as picking person B and then person A. Because the winners can be arranged in two ways, there are two times as many permutations as we need. Thus we need to divide the number of permutations by two.

More formally, each combination of k items can be written $k!$ ways. So we need to divide out $k!$ from our expression for permutations. Thus the number of combinations of length k from a set of size n is $C(n,k) = \frac{n!}{(n-k)!k!}$. Notice this is just $P(n,k)$ with $k!$ divided out (because each combination can be written in $k!$ ways).

We often write $C(n,k)$ as $\binom{n}{x}$ and pronounce this "n choose k".

7 Binomial distribution

7.1 Probability of a sequence

Say you have a sequence of independent binary outcomes that take the value 1 with probability p. We can compute the probability of a sequence of n such binary outcomes by multiplying the probability of each independent outcome. Note that if a binary outcome is 1 with probability p then it is 0 with probability $1-p$, as probability distributions must sum to 1.

Example 7.1. If a team wins a game with probability .7 then they will lose with probability .3, because probability distributions must sum to 1.

Example 7.2. A vendor gets a good review with probability p. If the vendor gets two good reviews and then a bad review, the probability of this sequence is $p * p * (1-p)$.

More generally, if you observe a sequence of n binary outcomes which take the value 1 with probability p, and you observe a total of k outcomes in the sequence that are equal to 1, then the probability of the sequence will be $p^k(1-p)^{n-k}$. This is because the probability of getting k ones is p^k and the probability of getting $n-k$ zeros is $(1-p)^{n-k}$. Note that if you observe k 1s in n trials then you must also observe $n-k$ zeros (because there are a total of n observations).

7.2 Number of ways to draw k 1s

Example 7.3. A vendor gets two good reviews and one bad review. Let 1 denote a good review and 0 denote a bad review. Then there are three possible ways for the vendor to get two good reviews: 110, 101, 011.

More broadly, there are $\binom{n}{k}$ possible ways to draw k 1s out of a sequence of n observations or trials (see Section 6.2 for more on this notation). One way to see why this is true is to imagine that each trial is a slot, so we have a total of n slots. If we observe k successes in n trials, we have to select k of the slots to fill with 1s. There are exactly $\binom{n}{k}$ possible ways to choose k 1s from n slots (Figure 7), in the same way that there $\binom{n}{k}$ possible ways to choose k members of a committee from a pool of n candidates. A few notes:

- Note that the probability of k successes is the same regardless of which k slots are assigned successes (i.e. order does not matter).

- Note also that if we draw k 1s (from any slots) the remaining $n-k$ slots will have to be 0s.

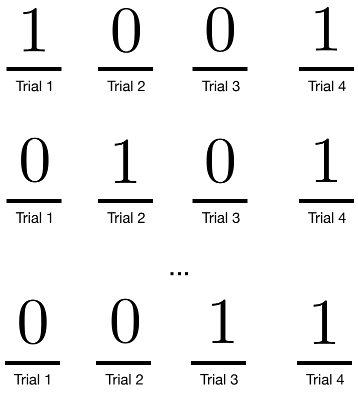


Figure 7: A few possible ways to draw $k = 2$ successes from $n = 4$ trials. In total, there are $\binom{n}{k}$ possible ways to draw k 1s from 4 slots. We only show 3 of the $\binom{n}{k}$ possibilities in this figure.

7.3 The binomial distribution

We are now ready to define the binomial distribution. Let X be a r.v. which counts the number of successes in n trials. Then the probability of k successes in n trials is $p(X = k) = \binom{n}{k}p^k(1-p)^{n-k}$, because the probability of each sequence with k successes is $p^k(1-p)^{n-k}$ and there are $\binom{n}{k}$ such sequences. This equation assigns some probability mass to all values of k from 0 to n. Distributing probability mass in this way is called the **binomial distribution**.

8 Maximum likelihood estimation

In problems in class, we sometimes assume that we know a r.v. takes some value with some probability. For instance, in Section 7.3 we assume we know that some binary r.v. takes the value 1 with probability p. In the real world, we usually don't actually know the actual parameters of a distribution; we have to "learn" them (see Section 5.4). We often use the letter theta θ to denote the parameters we want to learn.

One major use of Bayes rule[1] is to estimate the probability of different parameters, given the data. The idea is that we observe the data $\mathcal{D}$ and we want to infer the possible parameters θ which may have generated the data. We can do this with Bayes rule as follows

[1] In machine learning at least; Bayes rule is used in many contexts.

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} \tag{9}$$

where we use $\mathcal{D}$ to denote the data and θ to denote the parameters.

Recall that $p(\mathcal{D}|\theta)$ is called the likelihood (Section 5.12). Because we can not observe θ, we need to make some guess or estimate about the value of the parameter, denoted $\hat{\theta}$ (the "hat" over the theta emphasizes that we are estimating the value of θ). What guess should we make? One option, is to choose the **maximum likelihood estimate**. This chooses the value $\hat{\theta}_{MLE}$ that makes the likelihood of the observed data $\mathcal{D}$ the highest.

Example 8.1. A basketball player hits the first 3 of a total of 4 free throws in a game. Let θ be the true (and unknown) probability that the player will hit a free throw. Let $\hat{\theta}_1$ and $\hat{\theta}_{MLE}$ be two estimates of θ. $p(\mathcal{D}|\theta_1) = .2^3 * .8$ and $p(\mathcal{D}|\theta_{MLE}) = .75^3 * .25$, where $\mathcal{D} = 1110$. Because θ_{MLE} is chosen to maximize the likelihood, $p(\mathcal{D}|\theta_{MLE}) > p(\mathcal{D}|\theta_1)$.

8.1 MAP estimation

An alternative to the MLE is MAP estimation, which selects the $\hat{\theta}$ which maximizes both the likelihood and the prior. Because the value $\hat{\theta}_{MLE}$ may overfit the data, without the influence of the prior, some people interpret the prior as a kind of regularization. This is a little beyond the bounds of 2301.

9 Other resources

There are many resources that cover similar materials to 2301. It often helps to read the same concept from multiple perspectives. A few free textbooks that you might find helpful are listed below.

- https://cs.carleton.edu/faculty/dln/book/
- https://www.openintro.org/book/os/
- https://drive.google.com/file/d/1VmkAAGOYCTORq1wxSQqy255qLJjTNvBI/view