LING 473: Day 4

START THE RECORDING

Random Variables

Announcements

- Project 1
  - Due Thursday at 11:45pm
  - Questions?
- Assignment 2: Probability
  - Due August 8 at 4:30pm
  - Includes today's lecture
- August 8, 10
  - I will be at a conference (don't come to class)
  - Lectures will be recorded and made available at class time

Review

- Discrete probability spaces can be broken into mutually-exclusive, collectively-exhaustive individual events.
- Compositional events are made up of some number of these individual events:

$$P(A \cup B) = P(A) + P(B)$$

- Intersection defines two events occurring in the same trial:

$$P(A \cap B)$$

Review

- If two events are independent, then the likelihood of their intersection is equal to the product of their individual likelihoods.

$$P(A \cap B) = P(A)P(B)$$

  - Test for independence using this formula
- For non-mutually exclusive events
  - $P(A \cup B) = P(A) + P(B) - \mathrm{P}(\mathrm{A} \cap B)$

Review

- Conditional probabilities assume one event, calculate the likelihood of the other:

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

- Therefore:

$$P(A \cap B) = P(A|B)P(B)$$

Theory to Practice

- If we find a theoretical probability space that seems to correspond with some real-world phenomenon, this might help us predict something about the future occurrence of that phenomenon
- When we do this, we want to maintain probability's mathematical soundness
  - This will let us take full advantage of the methodology
- It so happens that probability spaces *are* useful for characterizing real-world phenomena, including NLP

Stochastic Trials

- Real-world events tend to have some degree of indeterminacy
  - Due to: failure to have all available information (theoretically impossible, see Heisenberg 1927)
  - Due to: failure to understand the system completely exactly
  - Due to: unforeseen events
- A measurable real-world process is called a stochastic process.
- A measurement, or observation, of a stochastic process is called a stochastic trial.

You may also encounter the alternate terms random process, or random trial. This is a very technical definition of *random*, however, as it is *characterizable*, but not deterministic; i.e., not “purely” random.

Random Variables

- Random variables bridge probability and statistics.
- Random variables equate a theoretical space with a measurable space of a stochastic process

Probability:

- Theoretically precise
- All outcomes are accounted for
- All outcomes are considered at once
- There is no trial, no event

Statistics:

- Empirical application
- We assume a well-formed probability space applies to a real-world phenomenon
- We predict future outcomes based on the model

Random Variables

A random variable is a function that maps a probability space Ω to the set of real numbers $\mathbb{R}$

$$X : \Omega \rightarrow \mathbb{R}$$

- Random variables allow us to…
  - use the machinery of probability to generalize over real-world events
  - describe the variability of stochastic trials
  - map outcomes to empirical, measurable values

Random Variables

- Random variables map every possible outcome in a sample space to a scalar (1-dimensional) value.
  - E.g., “2”, “0.15893”, “1,000,001”
  - Like events, we will use an upper case, italic letter.
  - For the sample space:

$$\Omega = \{ a, b, c, d, e, \ldots z\}$$

we can define the random variable

$$W = the\ number\ of\ times\ the\ letter\ appears\ in\ the\ document$$

Random Variables

- Random variables are not events.
- An event partitions a sample space into two subsets *E* and E^C
  - E.g., “the letter is a vowel” = { a, e, i, o, u }
  - Yes or no.
- A random variable maps a sample space into singular values.
  - E.g., “the number of vowels in the document” = {0,1,2, … }
  - Some range of numerical values

Discrete vs Continuous

- Like events, random variables can be discrete or continuous
  - Discrete random variables assume a finite set of values
  - However, if it is a count (e.g., countably infinite, 0, 1, 2, …) it isn't necessarily bounded, but still considered discrete.
- We've mentioned continuous sample spaces but didn't discuss their events
  - Ω = { *the distance between planets* }
    E = { *the distance is 54,523,189 km* }
    P(E) = 0.0
  - Continuous variables will help here.

Discrete vs Continuous

X = { *the number of miles (to the nearest mile) a commuter drives to work* }

X is a <u>discrete random variable</u>

X = { *the distance a commuter drives to work* }

X is a <u>continuous random variable</u>

X = { *the commuter drives to work* }

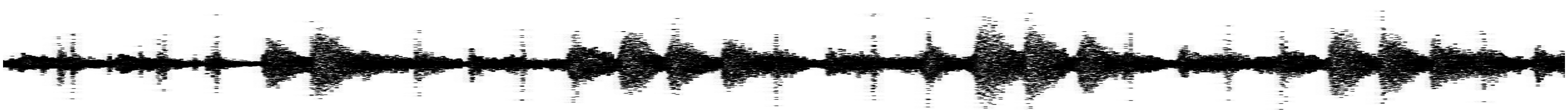
X is an <u>event</u>

Discrete Random Variables

- Random variables in computational linguistics are often counts
  - number of times a noun follows a determiner in a corpus
  - number of bytes downloaded form a URL
  - number of people in a study whose speech reflects a dialect feature
  - number of times a pair of words occur together in a corpus
  - number of times a word is used as a verb

Continuous Random Variable Examples

- More common in speech settings
  - duration of a phonological segment in a speech corpus
  - discourse particle usage interval timing in a sample of recorded discourse
  - F1 (1st formant) value in phonetic analysis
  - average frequency in voice recognition



Defining Discrete Random Variables

- Sometimes, outcomes in the sample space suggest certain numeric values for the random variable
- For example, rolling a six-sided die

Defining Discrete Random Variables

- The easiest way to define discrete random variables is to use the numeric value that it measures.

  X = { *the value of the roll of a single die* }

  "1"→ 1, "2" → 2, "3" → 3, "4" → 4, "5" → 5, "6" → 6

- or:

$$X = \begin{cases} 1, if\ the\ die\ shows\ 1, \\ 2, if\ the\ die\ shows\ 2, \\ 3, if\ the\ die\ shows\ 3, \\ 4, if\ the\ die\ shows\ 4, \\ 5, if\ the\ die\ shows\ 5, \\ 6, if\ the\ die\ shows\ 6. \end{cases}$$

Defining Discrete Random Variables

- The random variable is defined in terms of six mutually-exclusive events (the side of the die facing up)
  Q: Do they have to be collectively exhaustive?
  A: Yes, a discrete random variable must define a value for every possible outcome

- How do we deal with unforeseen outcomes (e.g., not dice-rolling?)
  - Just define a value that means “unobserved outcome”

Defining Discrete Random Variables

- Q: What about this: $X = \begin{cases} 88.8, if\ the\ die\ shows\ 1, \\ -2, if\ the\ die\ shows\ 2, \\ 123, if\ the\ die\ shows\ 3, \\ 0, if\ the\ die\ shows\ 4\ or\ 5, \\ 6.02 \times 10^{23}, otherwise. \end{cases}$?

- A: Ok. The values of a discrete random variable are arbitrary.

We're really only interested in the random variable's probabilities, which are defined in terms of their values. The numeric values that a random variable takes on only establish a *correspondence* between some event and the probability of that event.

Defining Discrete Random Variables

- We can even use text
  - This random variable captures whether a roulette wheel comes up red or black

$$W = \begin{cases} red \\ black \end{cases}$$

  - This random variable will be used to model the traditional gender categories

$$X = \begin{cases} male \\ female \end{cases}$$

Defining Continuous Random Variables

- For continuous random variables, we obviously cannot list a value for every point in the range
- Usually, the variable is defined as the real-valued data observation

- Continuous random variables can also be defined according to a continuous function, but this is less useful for modeling *measurements*

Events vs Random Variables

- An event is a single outcome, or some subset of outcomes, from Ω

$$\text{"}\textit{the total showing on the two dice is seven}\text{"}$$
$$E = \{\ (1{,}6), (2{,}5), (3{,}4), (4{,}3), (5{,}2), (6{,}1)\}$$

- A random variable is a function that maps any possible outcome to a real number (or quantifiable label)

$$X = F(\textit{the total showing on the two dice in a trial})$$

Event or Random Variable?

- Average F1 formant value in the recording
- The F1 value is above 600MHz
- The coin shows 'heads' ten times in a row
- "Call me Ishmael" are the first three words in the book
- The number of words before the word "Ishmael" in the book
- The number of clinical trial studies not discussing β-adrenergic blocking agents

Probability and Random Variables

- Random variables on their own are not all that useful
- A random variable is just a function that maps an outcome to some real number, it says nothing about the *likelihood* of getting that value
- We have seen the use of counting to get the probability for each outcome in the sample space. E.g.,

$$P(A) = \frac{|A|}{|\Omega|}$$

- For discrete spaces, we can calculate probability the same way.

Probability and Random Variables

- Random variables give us more tools beyond counting
- We introduce the idea of a **probability distribution**

A probability distribution is a function that maps all possible values (for discrete random variables) or ranges of values (for continuous random variables) of a random variable into a well-formed ("proper") probability space

Probability Distributions

The probability that the discrete random variable X will have the value x is notated by

$$P(X = x)$$

Alternate notation: $\rho_X(x)$

This is the probability mass function (pmf) of X

The probability that the continuous random variable X will have a value between a and b is notated by

$$P(a \leq X \leq b) = \int_a^b f_X(x)dx$$

The function $f_X(x)$ is the probability density function (pdf) of X

Discrete Random Variables

- For discrete random variables, we have to list the probability for each possible value.

use the lower case letter of the random variable to signify a single trial when defining discrete random variable probabilities

$$P(X = x) = \begin{cases} 0.1667, if\ x = 1; \\ 0.1667, if\ x = 2; \\ 0.1667, if\ x = 3; \\ 0.1667, if\ x = 4; \\ 0.1667, if\ x = 5; \\ 0.1667, if\ x = 6; \\ 0, otherwise. \end{cases}$$

the last line can be left off

- This is why the values for *x* are arbitrary. We can choose it to most conveniently match the empirical value (like the number on a die side)

Discrete Random Variables

$$P(X=x)=\begin{cases} 0.1667, if\ x=1; \\ 0.1667, if\ x=2; \\ 0.1667, if\ x=3; \\ 0.1667, if\ x=4; \\ 0.1667, if\ x=5; \\ 0.1667, if\ x=6; \\ 0, otherwise. \end{cases} \longleftrightarrow X=\begin{cases} 1, if\ the\ die\ shows\ 1, \\ 2, if\ the\ die\ shows\ 2, \\ 3, if\ the\ die\ shows\ 3, \\ 4, if\ the\ die\ shows\ 4, \\ 5, if\ the\ die\ shows\ 5, \\ 6, if\ the\ die\ shows\ 6. \end{cases}$$

Probability Mass Function

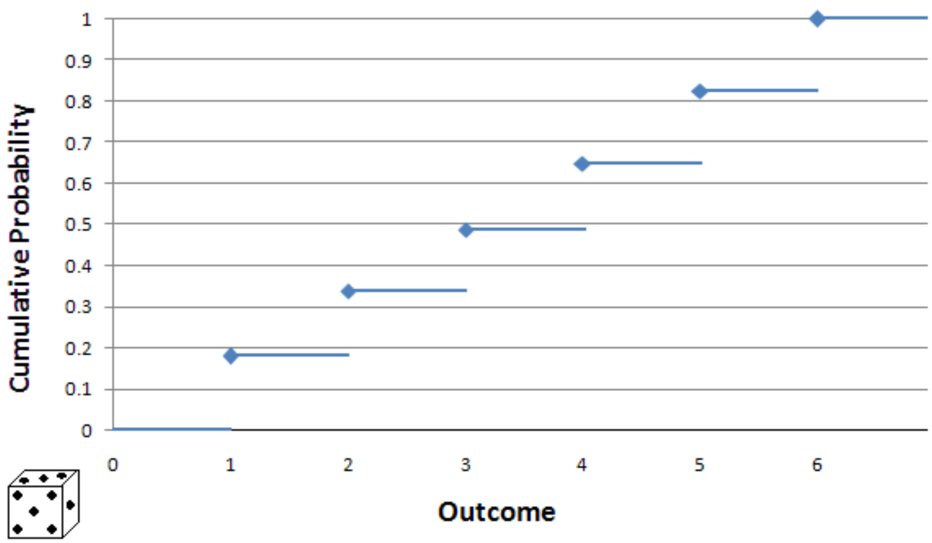
- The probability $P(X = x)$ of a discrete random variable is called its probability mass function (pmf)
- This doesn't work for continuous random variable Y because $\forall\, x, P(X = x) = 0$

Cumulative Distribution Function

- The cumulative distribution function (cdf) shows the accumulated mass of the pmf

  $P(X \leq x)$

- For a discrete random variable, this will be a step function



Probability Density Function

- We need to use some calculus to describe continuous random variable X. The first step is to find a cumulative distribution function

$$P(X \leq x) = \int_{-\infty}^{x} f_X(u)du$$

Then, the derivative of this, $f_X(x)$ is the pdf of the continuous random variable

$$\mathrm{pdf}_X = f_X(x) = \frac{d \int_{-\infty}^{x} f_X(u)du}{dx}$$

Probability Density Function

$$\mathrm{pdf}_X = f_X(x) = \frac{d\int_{-\infty}^{x} f_X(u)du}{dx}$$

- We cannot use $P(X = x)$ for the right hand side. Why?
  - In a continuous function, the likelihood of any single point is zero.
  - It is conventional to use $f_X(x)$ for the pdf.
  - (Example forthcoming)

Where do the probabilities come from?

- They mean something only "by construction"
- For discrete event *E*, we conjured probability $P(E)$
- For discrete random variable *X*, we conjured probability

  $P(X = x)$
- For continuous random variable *X*, we conjured cumulative distribution

$$P(X \leq x) = \int_{-\infty}^{x} f_X(u)du$$

Where do the probabilities come from?

Because a random variable—like an event—encapsulates all possible outcomes in a sample space, its probability function *meaningfully* characterizes that sample space

We use the random variable as an estimate—or proxy—for the entire sample space

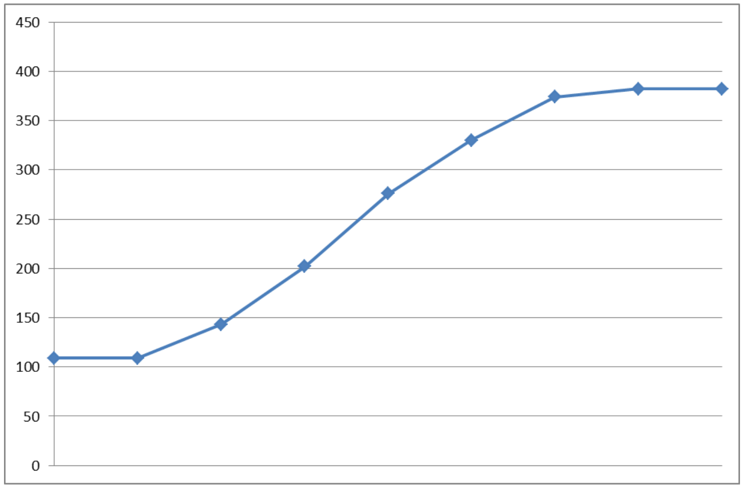
And, conversely, *use observed probabilities* to define the random variable. Let’s study a continuous random variable as an example.

Continuous random variable: example

X = { *rhyme duration of* /bay/ *time (ms.) in the test population* }

raw data: {202, 374, 279, 330, 382, 140, 109 }

sorted: { 109, 140, 202, 279, 330, 374, 382 }



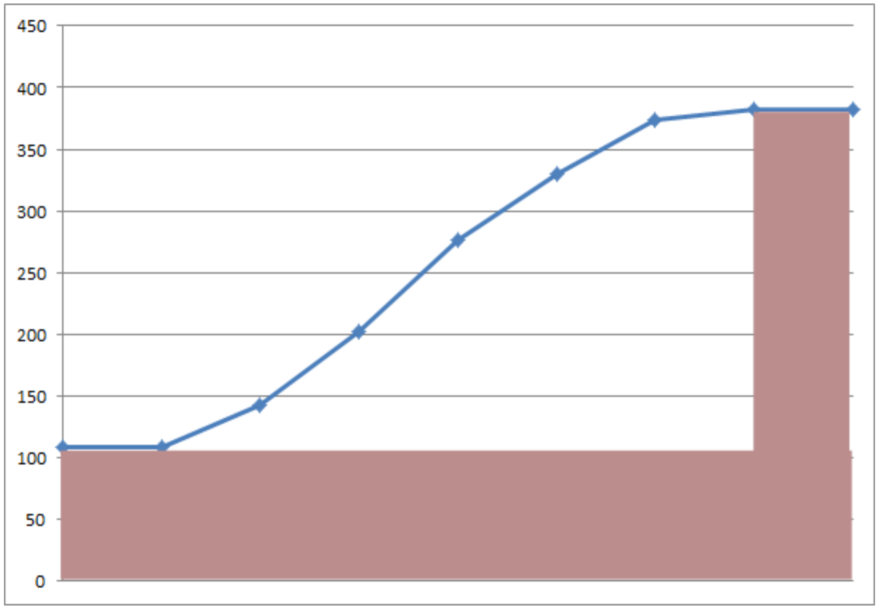
Here's our raw data

They are equally spaced because each observation is equally important

We extend the first and last points horizontally to show that there are no further observations

Continuous random variable: example

Normalize the data by first throwing away the shaded part…



We are working towards making our data into a cdf

We must normalize the data to meet the following cdf criteria:

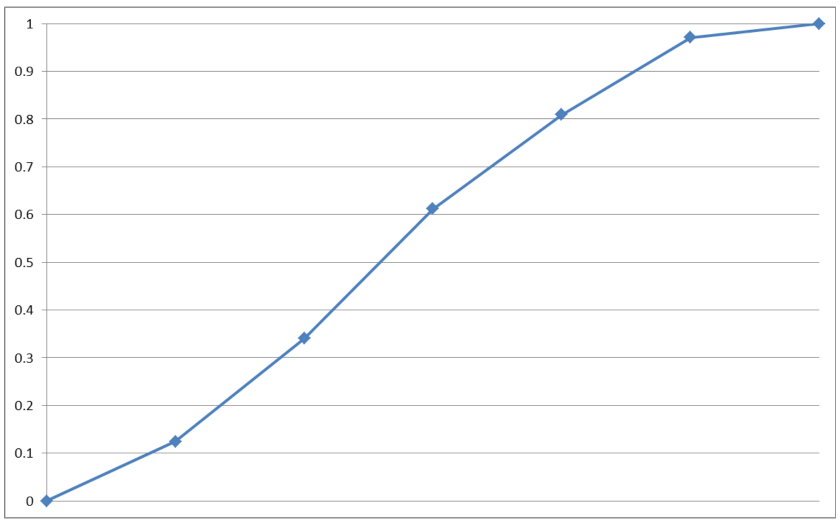
- It has the value 0 at $-\infty$
- It has the value 1 at $+\infty$
- It is monotonically non-decreasing

We also remember the normalization factors we used so that we can reverse this process

Continuous random variable: example

…and scaling the top value to 1

(this is the "by construction" part)



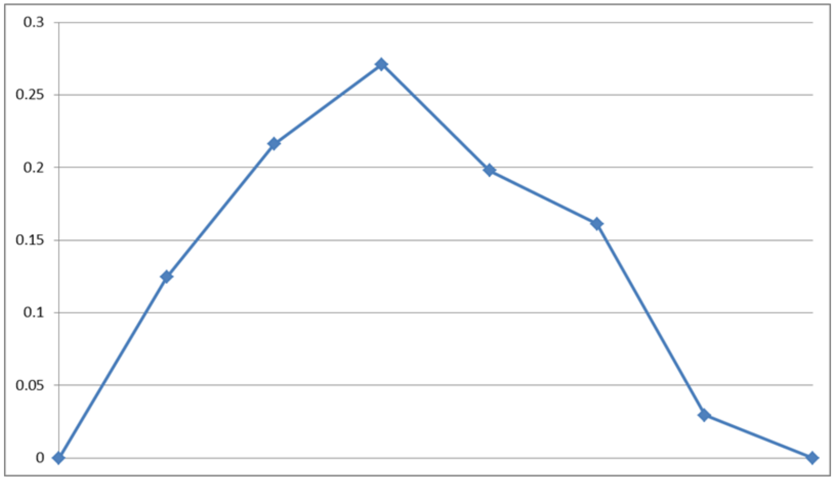
x

This now meets the definition of a cumulative density function

$$P(X \leq x) = \int_{-\infty}^{x} f_X(u)du$$

Continuous random variable: example

Finally, take the derivative. This is the set of differences between adjacent points in the cdf



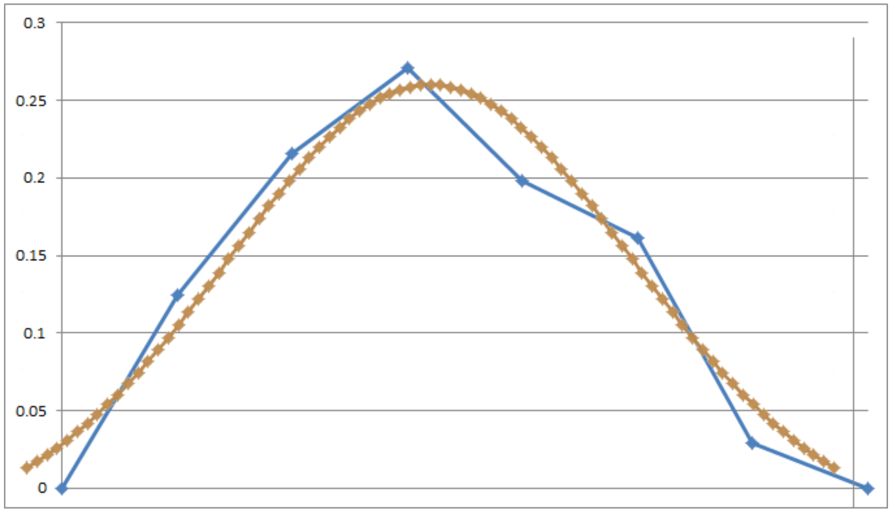
This is a probability density function (pdf) for our random variable

$$f_X(x) = \frac{d\int_{-\infty}^{x} f_X(u)du}{dx}$$

Because of the way we constructed it, the total area under this curve will equal 1.0.

Continuous random variable example

Even though we only had a few data points, it looks like our phonetics data fit the shape of a normal curve pretty closely



The normal curve is a type of probability distribution we'll be studying later

Summary of the example

- The example shows that we can take a few raw data points make a general statement about our data:

"Measurement of rhyme duration of the syllable /bay/ in the test population for speakers in our study is approximately normal."

- If we assume that this distribution *characterizes* our continuous random variable, then we can use this distribution to predict the studied feature in other speakers, or in the general population

Summary: pmf / pdf / cdf

The probability mass function (pmf) of a discrete random variable X is notated by

$$P_X\ (X = x)$$ alternate notation: $\rho_X(x)$

The probability density function (pdf) of a continuous random variable X is notated by

$$f_X(x)$$

For either type, the cumulative distribution function (cdf) is notated by

$$P_X(X \leq x)$$ alternate notation: $F_X(x)$

The subscripted random variable is usually omitted, so you have to remember that P is a different function for each random variable that you're working with

Conditional Independence

- Two events A and B are conditionally independent given a third event K if they are independent in their conditional probability distributions:

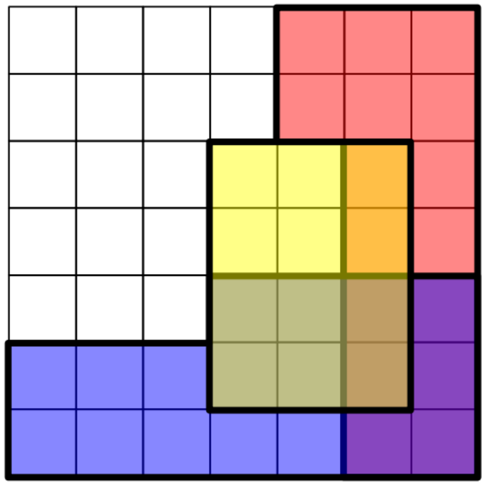
$$P(A|B \cap K) \stackrel{?}{=} P(A|K)$$
$$P(B|A \cap K) \stackrel{?}{=} P(B|K)$$

- Put another way:

$$P(A \cap B|K) \stackrel{?}{=} P(A|K)P(B|K)$$

(Note that | has lowest precedence)

Conditional independence



$$P(R) = \frac{16}{49}$$

$$P(B) = \frac{18}{49}$$

$$P(R \cap B) = \frac{6}{49} \neq P(B) \times P(R)$$

none of these are independent

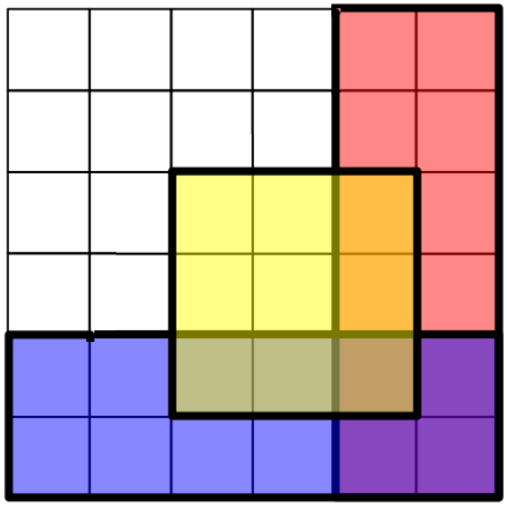
… but they are conditionally independent given Y

$$P(R|Y) = \frac{4}{12} = \frac{1}{3}$$

$$P(B|Y) = \frac{6}{12} = \frac{1}{2}$$

$$P(R \cap B|Y) = \frac{2}{12} = \frac{1}{6} = P(R|Y) \times P(B|Y)$$

Conditional independence



$$P(R) = \frac{12}{36} = \frac{1}{3}$$

$$P(B) = \frac{12}{36} = \frac{1}{3}$$

$$P(R \cap B) = \frac{4}{36} = \frac{1}{9} = P(R)P(B)$$

R and B are independent

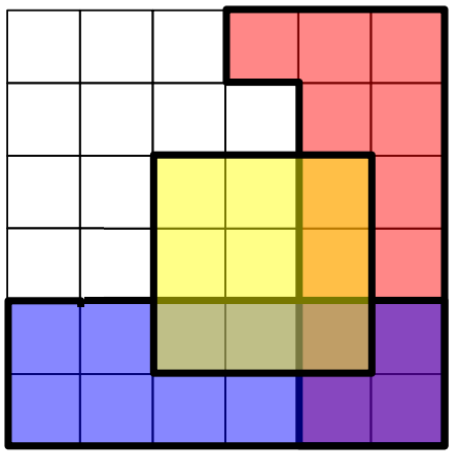
$$P(R|Y) = \frac{3}{9} = \frac{1}{3}$$

$$P(B|Y) = \frac{3}{9} = \frac{1}{3}$$

$$P(R \cap B|Y) = \frac{1}{9} = P(R|Y)P(B|Y)$$

R and B are {also, still} conditionally independent given Y

Conditional independence



$$P(R) = \frac{13}{36}$$

$$P(B) = \frac{12}{36} = \frac{1}{3}$$

$$P(R \cap B) = \frac{4}{36} = .1111$$

$$P(R)P(B) = \frac{13}{108} = .1214$$

…these are not equal, so R and B are dependent. But…

$$P(R|Y) = \frac{3}{9} = \frac{1}{3}$$

$$P(B|Y) = \frac{3}{9} = \frac{1}{3}$$

$$P(R \cap B|Y) = \frac{1}{9}$$

R and B can be *conditionally* independent given Y, even if they are dependent in the absence of information about Y