Probability Refresher

Sample Space of an Experiment

- A random experiment is one whose outcome cannot be predicted with certainty.
  e.g. Toss of a coin, roll of a dice, etc.

- **Sample space** (S) represents the set of all possible outcomes of the experiment.

  e.g. Toss of a coin = {H, T}
  Roll of a single dice = {1, 2, 3, 4, 5, 6}
  Roll of two dice = {11, 12, …, 16,
  21, 22, ..., 26,
  ...,
  61, 62,..., 66 }

Sample Space and Event

- An event (E) is a subset of S (sample space).
- Example:
  - Tossing a coin:
    S = {H, T}   E = {H} is an event.
  - Tossing a coin twice:
    S = {HH, HT, TH, TT}
    E = {HH, HT} is an event which describes the set which has the first outcome as heads
  - Tossing a dice:
    S = {1, 2, 3, 4, 5, 6}
    E = {2, 4, 6} is an event which describes the set in which the number is even.

Random Variable

- Suppose to each outcome in the sample space, we associate a value.
  e.g. for heads, +1 and for tails -1
  for even rolls of dice +1 and for odd rolls of dice -1

- Such a variable is known as a random variable.
- Formal definition:
  A **random variable** is a function that assigns a real number to each outcome in the sample space of a random experiment.
- It is generally denoted by X.
  If X takes a value x, it is written as: X = x

Random Variable

- If we restrict the random variable to the Boolean set, it may be defined as a function f:

  f is a function defined over S as follows:

$$f: S \rightarrow \{0, 1\}$$

f maps every value in S to either 0 (failure) or 1 (success)

Examples of Random Variables:

- Suppose each experiment is tossing of 4 coins. You do this experiment multiple times
  Results:
  {HHHT}
  {HTHT}
  …
  If you count the number of heads in each experiment, it would be a random variable (X)
  X could have values from 0 to 4.
- Each element of the sample space would map to one value of X

Examples of Random Variables:

- Suppose each experiment is sampling 100 people from a population and measure their heights.

  If you measure the average of the heights of the 100 samples, you would get a random variable (X). It would be a continuous variable.

  Each element of the sample space would map to a value of X

Random Variable

- Each random event A has a **probability P(A)** associated with it.
- It defines the fraction of the sample space in which A is true.

$$\mathrm{P(A)} = \frac{\text{Number of events in which A is true}}{\text{Total number of events in } S}$$

e.g. in case of toss of a dice, probability of even number:

$$P(A) = \frac{3}{6} = 0.5$$

Useful Theorem

- 0 <= P(A) <= 1, P(True) = 1, P(False) = 0,
  P(A or B) = P(A) + P(B) - P(A and B)

➔ P(A) = P(A ^ B) + P(A ^ ~B)

A = [A and (B or ~B)] = [(A and B) or (A and ~B)]

P(A) = P(A and B) + P(A and ~B) – P((A and B) and (A and ~B))

$P(A) = P(A \text{ and } B) + P(A \text{ and } {\sim}B) - P(\cancel{A \text{ and } B \text{ and } A \text{ and } {\sim}B})$

Definition of Conditional Probability

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

Corollary: The Chain Rule

$$P(A \wedge B) = P(A|B)\, P(B)$$

$$P(C \wedge A \wedge B) = P(C|A \wedge B)\, P(A|B)\, P(B)$$

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$ Bayes' rule

we call P(A) the "prior"

and P(A|B) the "posterior"

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London,* **53:370-418**

…by no means merely a curious speculation in the doctrine of chances, but necessary to be solved in order to a sure foundation for all our reasonings concerning past facts, and what is likely to be hereafter…. necessary to be considered by any that would give a clear account of the strength of *analogical* or *inductive reasoning*…

Applying Bayes Rule

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

A = you have the flu,   B = you just coughed

Assume:
P(A) = 0.05
P(B|A) = 0.80
P(B| ~A) = 0.2

what is P(flu | cough) = P(A|B)?

what does all this have to do with function approximation?

After all, that's what we are looking for

Function approximation by probability

- Our aim is to approximate the class separating function
  F: X -> Y
  X is the set of attributes and Y is the class.
- In case of probabilistic reasoning, we approximate it with the conditional probability
  P(Y | X)
  we find the probability of each class, given the data i.e.
  P(Y = 1| X) and P(Y = 0 | X).
- How do we find the probability of class given the data?
- Need a joint distribution

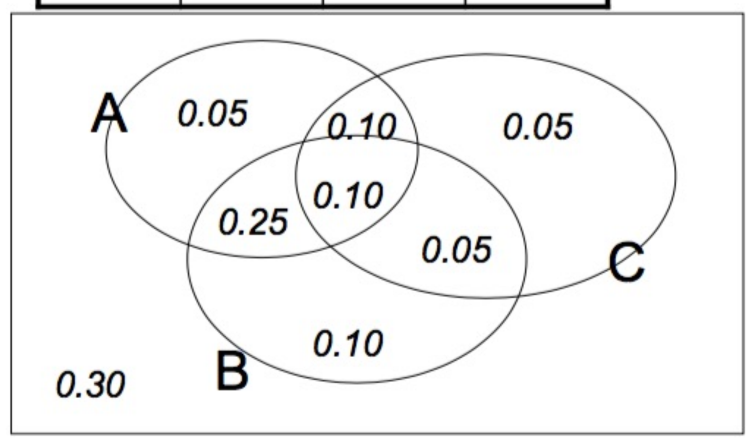
The Joint Distribution

Example: Boolean variables A, B, C

Recipe for making a joint distribution of M variables:

1. Make a truth table listing all combinations of values of your variables (if there are M Boolean variables then the table will have 2^M rows).
2. For each combination of values, say how probable it is.
3. If you subscribe to the axioms of probability, those numbers must sum to 1.

A	**B**	**C**	**Prob**
0	0	0	0.30
0	0	1	0.05
0	1	0	0.10
0	1	1	0.05
1	0	0	0.05
1	0	1	0.10
1	1	0	0.25
1	1	1	0.10



Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

One you have the JD you can ask for the probability of any logical expression involving your attribute

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor Male) = 0.4654

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Using the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

P(Poor) = 0.7604

$$P(E) = \sum_{\text{rows matching } E} P(\text{row})$$

Inference with the Joint

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

$$P(E_1 \mid E_2) = \frac{P(E_1 \wedge E_2)}{P(E_2)} = \frac{\sum_{\text{rows matching } E_1 \text{ and } E_2} P(\text{row})}{\sum_{\text{rows matching } E_2} P(\text{row})}$$

P(Male | Poor) = 0.4654 / 0.7604 = 0.612

Learning and the Joint Distribution

gender	hours_worked	wealth	
Female	v0:40.5-	poor	0.253122
		rich	0.0245895
	v1:40.5+	poor	0.0421768
		rich	0.0116293
Male	v0:40.5-	poor	0.331313
		rich	0.0971295
	v1:40.5+	poor	0.134106
		rich	0.105933

Suppose we want to learn the function f: <G, H> → W

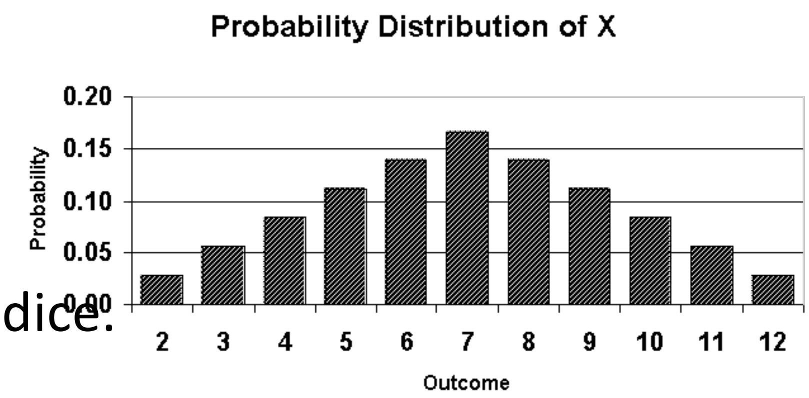
Equivalently, P(W | G, H)

Solution: learn joint distribution from data, calculate P(W | G, H)

e.g., P(W=rich | G = female, H = 40.5-) =

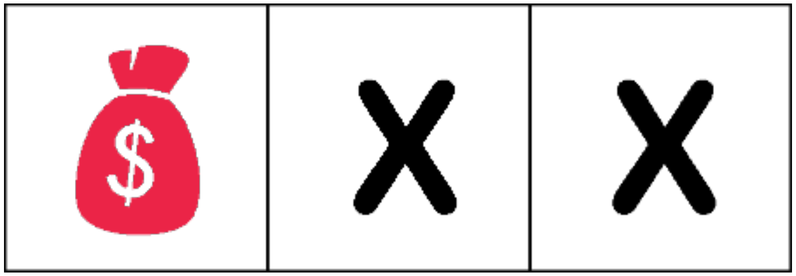
Let's get back to random variables

- Each random variable X has a domain ***D(x)*** associated with it. It is the set of outcome values possible.
  e.g. for a dice D(x) = {1, 2, 3, 4, 5, 6}
  for a Boolean outcome D(x) = {0, 1}.



- For a pair of dice, D(x) = {11, …., 66}.
  Suppose X is the sum of the outcome of the dice.
  X can range from 2 to 12.
  Each value of X has a different probability.
- A plot of the random variable and probability values is called the probability distribution plot and the function is called probability distribution function (pdf) or probability mass function (pmf).

Example:



- Imagine there are 3 doors – one of them has a treasure, the other 2 nothing.
- Each one is equally likely to be selected. Once you open a door, you don't open it again.
- Define random variable X as the number of doors needed to open before finding the treasure.
  X can have values {1, 2}. Find the probability distribution of X.

  P(X = 1) = 1/3
  P(X = 2) = (2/3) * (1/2)
  -- Assume a person is smart enough to figure it out after two attempts -- ☺

Example 2: Consider a group of five potential blood donors—a, b, c, d, and e—of whom only a and b have type O+ blood. Five blood samples, one from each individual, will be typed in random order until an O+ individual is identified. Let the rv Y = the number of typings necessary to identify an O+ individual. Then the pmf of Y is

$$p(1) = P(Y = 1) = P(a \text{ or } b \text{ typed first}) = \frac{2}{5} = .4$$

$$p(2) = P(Y = 2) = P(c, d, \text{ or } e \text{ first, and then } a \text{ or } b)$$

$$= P(c, d, \text{ or } e \text{ first}) \cdot P(a \text{ or } b \text{ next} \mid c, d, \text{ or } e \text{ first}) = \frac{3}{5} \cdot \frac{2}{4} = .3$$

$$p(3) = P(Y = 3) = P(c, d, \text{ or } e \text{ first and second, and then } a \text{ or } b)$$

$$= \left(\frac{3}{5}\right)\left(\frac{2}{4}\right)\left(\frac{2}{3}\right) = .2$$

$$p(4) = P(Y = 4) = P(c, d, \text{ and } e \text{ all done first}) = \left(\frac{3}{5}\right)\left(\frac{2}{4}\right)\left(\frac{1}{3}\right) = .1$$

$$p(y) = 0 \quad \text{if } y \neq 1, 2, 3, 4$$

Binary Variables

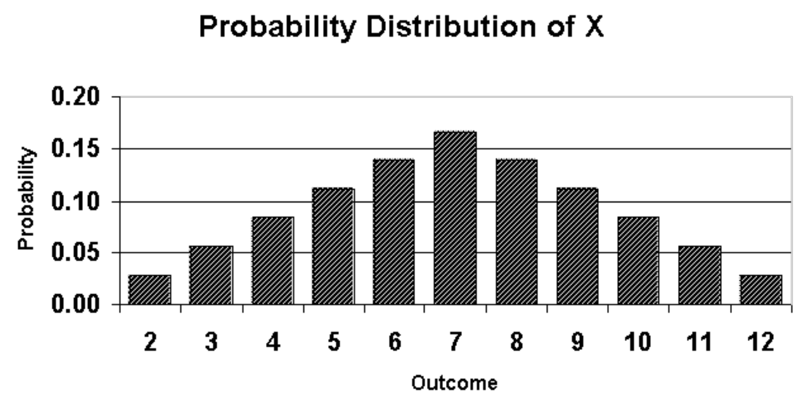
- Let's focus on the case where the random variable X can only take two values {0, 1}.
- X is said a Boolean random variable.
- For every outcome in the sample space, you associate 0 (failure) or 1 (success).
- Example:
  - Coin toss – Heads = 1, Tails = 0
  - Lottery - Winning Number = 1, Rest of the numbers = 0
  - …

Expected Value and Variance

- **Expected value** of a discrete random variable under P:

$$E_P(X) = \sum_{x \in D(x)} x\, P(x)$$

take each value and multiply by its probability



- **Variance** of the random variable under P:

$$var_P(X) = \sum_{x \in D} \left(x - E_P(x)\right)^2 P(x)$$

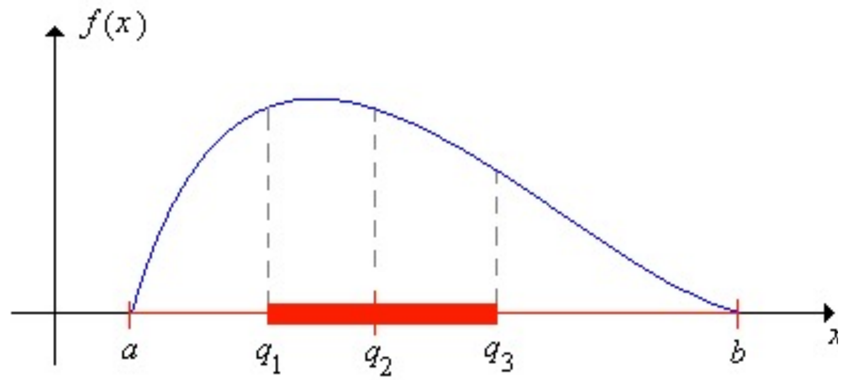
Shortcut formula:

$$var_P(X) = E_P(X^2) - [E_P(X)]^2$$

Expected Value and Variance

- **Expected value** of a continuous random variable under a continuous probability function f:

$$E_f(X) = \mu = \int_{-\infty}^{\infty} x f(x) dx$$



Continuous Probability Density Function

- **Variance** of the random variable under P:

$$var_f(X) = \int_{-\infty}^{\infty} (x - \mu)^2 f(x) dx$$

Shortcut formula:

$$var_f(X) = E_f(X^2) - \left[E_f(X)\right]^2$$

Example:

You draw one card from a standard deck of playing cards. If you pick a heart, you will win$10. If you pick a face card (K, Q, J), which is not a heart, you win $8. If you pick any other card, you lose $6. Do you want to play? Explain.

Solution

Let X be the random variable that takes on the values 10, 8 and –6, the values of the winnings. First, we calculate the following probabilities:

$$P(X=10)=\frac{13}{52},\ P(X=8)=\frac{9}{52},\text{ and }P(X=-6)=\frac{30}{52}.$$

The expected value of the game is

$$\begin{aligned}E(X)&=P(X=10)*10+P(X=8)*8-P(X=-6)*6\\&=\frac{13}{52}*10+\frac{9}{52}*8-\frac{30}{52}*6\\&=\frac{130+72-180}{52}\\&=\frac{22}{52}\end{aligned}$$

Since the expected value of the game is approximately \$.42, it is to the player's advantage to play the game.

Binary Variables

- Consider coin flipping which has 2 outcomes (heads = 1 and tails = 0)
  If you know probability of heads, you know the distribution.
  Let's say
  $$p(x = 1|\ \mu) = \mu \qquad p(x = 0|\ \mu) = 1 - \mu$$
- Probability of heads in **one** coin flip makes the Bernoulli Distribution (Bern):

$$Bern(x \mid \mu) = \mu^{x}(1 - \mu)^{1-x}$$

Not convinced?
Plug in x = 0 and x = 1 in this equation to see

Easy to show that

$$E_{Bern}(x) = \mu$$

$$var_{Bern}(x) = \mu(1 - \mu)$$

What do you notice?
If I know μ, I know everything about the distribution.

Binomial Distribution

- Now imagine you throw that same coin **N times** and you want to find the probability of **m heads** where m can be from 0 to N.

$$p(m\ heads|\ N, \mu)$$

- This is called the Binomial distribution.

$$Bin(m \mid N, \mu) = \binom{N}{m} \mu^m (1-\mu)^{N-m}$$

It is easy to show that:

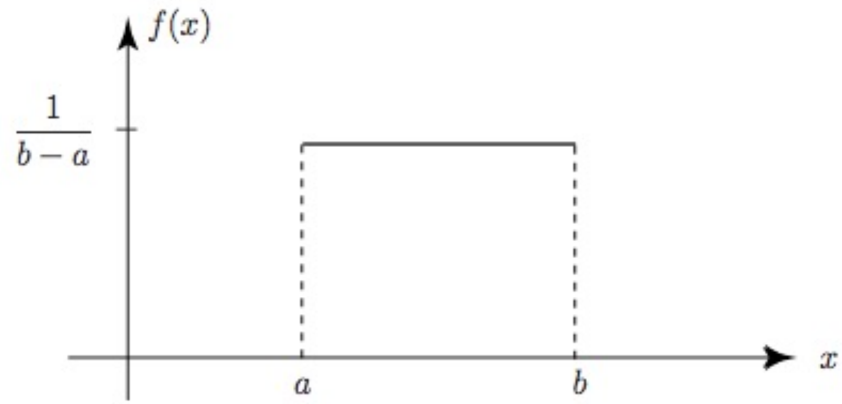
$$E[m] = \sum_{m=1}^{N} m\ Bin(m \mid N, \mu) = N\ \mu$$

$$var[m] = \sum_{m=1}^{N} (m - E[m])^2\ Bin(m \mid N, \mu) = N\ \mu(1-\mu)$$

What do you notice?
If I know N and μ, I know everything about the distribution.

Uniform Probability Distribution

- Continuous Distribution:
  Density plot shown on right



The function $f(x)$ is defined by:

$$f(x) = \begin{cases} \frac{1}{b-a}, & a \le x \le b \\ 0 & \text{otherwise} \end{cases}$$

$$\begin{aligned} E(X) &= \int_{-\infty}^{\infty} x f(x) \mathrm{d}x = \int_a^b x \frac{1}{b-a} \mathrm{d}x = \frac{1}{2(b-a)} \left[x^2\right]_a^b \\ &= \frac{b^2 - a^2}{2(b-a)} \\ &= \frac{b+a}{2} \end{aligned}$$

What do you observe?
Just need a and b to know everything about the distribution.

$$\begin{aligned} V(X) &= E(X^2) - [E(X)]^2 \\ &= \int_a^b x^2 . \frac{1}{b-a} \mathrm{d}x - \left(\frac{b+a}{2}\right)^2 = \frac{1}{3(b-a)} \left[x^3\right]_a^b - \left(\frac{b+a}{2}\right)^2 \\ &= \frac{b^3 - a^3}{3(b-a)} - \left(\frac{b+a}{2}\right)^2 \\ &= \frac{b^2 + ab + a^2}{3} - \frac{b^2 + 2ab + a^2}{4} \\ &= \frac{(b-a)^2}{12} \end{aligned}$$