September 28, 2017

CS 361: Probability & Statistics

Random variables

Chebyshev's inequality

For any random variable X and value a

$$P(\{|X - E[X]| \geq a\}) \leq \frac{\mathrm{var}[X]}{a^2}$$

Or if we let σ be the standard deviation of X and substitute $a = k\sigma$ we have

$$P(\{|X - E[X]| \geq k\sigma\}) \leq \frac{1}{k^2}$$

The probability that X is greater than k standard deviations from the mean is small. Look familiar?

Proving Chebyshev

Proving

$$P(\{|X - E[X]| \geq a\}) \leq \frac{\text{var}[X]}{a^2}$$

Write U for the random variable $(X - E[X])^2$

Markov's inequality tells us then that for any w, we will have

$$P(\{|U| \geq w\}) \leq \frac{E[|U|]}{w}$$

Using $w = a^2$ we have

$$P(\{|U| \geq w\}) = P(\{|X - E[X]| \geq a\}$$

Also from the definition of U, w, and variance we have

$$\frac{E[|U|]}{w} = \frac{\text{var}[X]}{a^2}$$

So substituting we get, as desired

$$P(\{|X - E[X]| \geq a\}) \leq \frac{\text{var}[X]}{a^2}$$

Sampling

So far we have been working on a number of problems where we suppose ahead of time that we know the distributions of the outcomes and hence the random variables in our experiments.

We eventually want to get to a place where we don't make that assumption and we guess what the distribution is after observing some runs of an experiment

In this context and others we will refer to our observations of experiments as **trials** or **samples** from the underlying distribution

Sampling: IID

If we have a set of data items x_i meeting the following:

a) they are independent

b) they were generated by the same process

c) the histogram of a very large set of the items looks increasingly like the probability distribution P(X) of some random variable X

We call this set of data items **independent, identically distributed samples of P(X)** or IID for short

Expectation of iid samples

Assume we have a set of N iid samples of a probability distribution P(X)

$$X_N = \frac{\sum_{i=1}^{N} x_i}{N}$$

X_N is a random variable, and if we take the expectation of both sides we get

$$E[X_N] = E\left[\frac{\sum_{i=1}^{N} x_i}{N}\right]$$

Using linearity, we get

$$E[X_N] = \frac{1}{N}\sum_{i=1}^{N} E[x_i]$$

But since x_i is a sample drawn from X

$$E[X_N] = \frac{1}{N}\sum_{i=1}^{N} E[X]$$

Simplifying we get

$$E[X_N] = E[X]$$

Variance of iid samples

Assume that X has a finite variance given by σ^2 . Let's find the variance of X_N

$$\text{var}[X_N] = \text{var}\left[\frac{\sum_{i=1}^{N} x_i}{N}\right]$$

Recall the property of variance

$\text{var}[kX] = k^2\text{var}[X]$ to get

$$\text{var}[X_N] = \frac{1}{N^2}\text{var}\left[\sum_{i=1}^{N} x_i\right]$$

Since the x_i are drawn independently from X we can use the fact that the variance of a sum of independent variables can be broken up into a sum of variances

$$\text{var}[X_N] = \frac{1}{N^2}\sum_{i=1}^{N}\text{var}[x_i]$$

Substituting

$$\text{var}[X_N] = \frac{1}{N^2}\sum_{i=1}^{N}\sigma^2$$

And simplifying, we get

$$\text{var}[X_N] = \frac{\sigma^2}{N}$$

Weak law of large numbers

With $X_N = \frac{\sum_{i=1}^{N} x_i}{N}$

The **weak law of large numbers** states that, if $P(X)$ has finite variance, then for any positive number ϵ

$$\lim_{N \to \infty} P(\{|X_N - \mathbb{E}[X]| \geq \epsilon\}) = 0.$$

Equivalently, we have

$$\lim_{N \to \infty} P(\{|X_N - \mathbb{E}[X]| < \epsilon\}) = 1.$$

Each form means that, for a large enough set of IID samples, the average of the samples (i.e. X_N) will, with high probability, be very close to the expectation $\mathbb{E}[X]$.

Proving the weak law

Proving

$$\lim_{N\to\infty} P(\{|X_n - E[X]| \geq \epsilon\}) = 0$$

Let's put X_N into Chebyshev's inequality

$$P(\{|X_N - E[X_N]| \geq \epsilon\}) \leq \frac{\text{var}[X_N]}{\epsilon^2}$$

Using what we showed a few slides ago $E[X_N] = E[X]$ we substitute to get

$$P(\{|X_N - E[X]| \geq \epsilon\}) \leq \frac{\text{var}[X_N]}{\epsilon^2}$$

Using $\text{var}[X_N] = \frac{\sigma^2}{N}$ we get

$$P(\{|X_N - E[X]| \geq \epsilon\}) \leq \frac{\sigma^2}{N\epsilon^2}$$

Taking the limit as N goes to infinity proves the weak law, the right hand side goes to 0, as desired

Who cares?

- The weak law tells us that if we observe enough data we can learn things about random variables
- It means we can do simulations instead of calculations in some cases
- It is useful for solving problems

Example: Roulette

A roulette wheel has 36 numbers, 18 are red and 18 are black. It also has a number of colorless zeros. If you bet 1 on red and a red number comes up, you receive a payout of 1, if a red number does not come up you lose 1.

On a wheel with 1 zero, what is the expected payout?

$$(1)P(\text{win}) - (1)P(\text{lose})$$

$$18/37 - 19/37 = -1/37$$

On a wheel with 2 zeros, what is the expected payout?

$$(1)P(\text{win}) - (1)P(\text{lose})$$

$$18/38 - 20/38 = -1/19$$

On a wheel with 3 zeros, what is the expected payout?

$$(1)P(\text{win}) - (1)P(\text{lose})$$

$$18/39 - 21/39 = -1/13$$

Fair games

- In some sense, a game isn't really "fair" unless it has an expected payout of 0
- Taking on a game with negative expected value means on average you will lose and the more times you play the more you will lose
- The weak law says a casino will be making about 5 cents per bet on red in roulette—now matter how much people bet or how lucky they're feeling

Example: birthdays

P1 and P2 are going to play a game where P1 bets that if 3 people are stopped on the street, they will have birthdays whose days of the week are in succession. If this happens P2 gives P1 $100, if not P1 gives P2 $1. What is the expected payout to P1?

The payout will be
100p - (1-p)1

Recall that we calculated p before and it was 1/49

So P1's payout is 52/49 or slightly more than a dollar

If P1 can find people willing to take this bet, they should do it all day

Example: ending a game early

Two players each put up $25 to play the following game: they will toss a fair coin. If it comes up heads, player H wins that toss and for tails, player T wins that toss. The first player to win 10 tosses gets all $50. One of the players has to leave when the game is scored 8-7 (H-T), how should the $50 be divided between them?

Either person could win if play were allowed to continue. We should divide up the $50 based on the expected earnings of each player at the current point in the game. Player H's expected value is

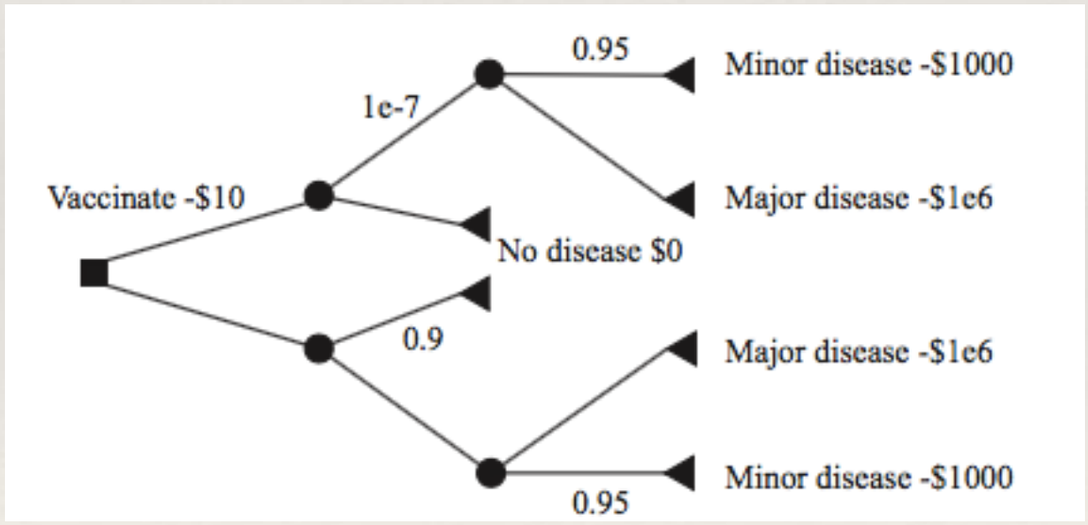
$$50\ P(\{\text{H wins from score 8-7}\}) + 0\ P(\{\text{T wins from score of 8-7}\})$$

H could win 10-7, 10-8, or 10-9. These are disjoint events

$P(\text{“10-7”}) = 1/4$	$P(\text{“10-8”}) = 2/8$	$P(\text{“10-9”}) = 3/16$	$E[H] = 50*11/16$
HH	HTH, THH	HTTH, THTH, TTHH	$= \$34.375$

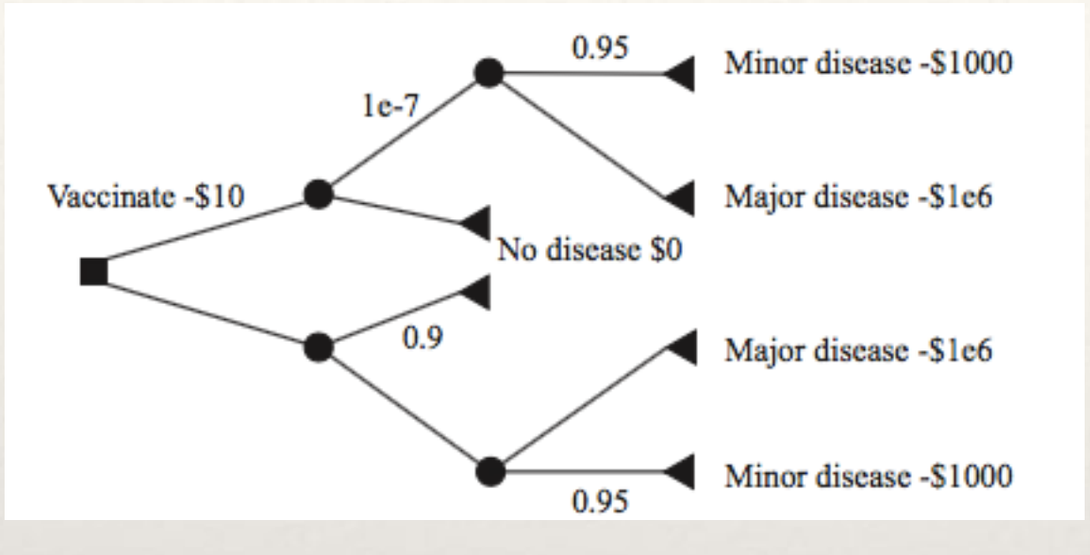
Example: decision tree

We have to decide whether to be vaccinated or not. It costs 10 to get the vaccine. If you're vaccinated you get the disease with probability 1e-7. If you're not you get it with probability 0.1 The disease is unpleasant and with probability 0.95 you will experience problems that cost you 1000 to have treated, but with probability 0.05 you get a major form of the disease and have to spend 1e6. Should you get vaccinated?



In a decision diagram, squares represent choices, circles random events, and triangles outcomes

Example: decision tree



What to compute first?

Expected cost of disease

(0.95)(1000) + (0.05)(1e6) = $50,950

Cost to vaccinate

10 + (1e-7)(50,950) = $10.01

Cost not to

(0.1)(50950) = $5095

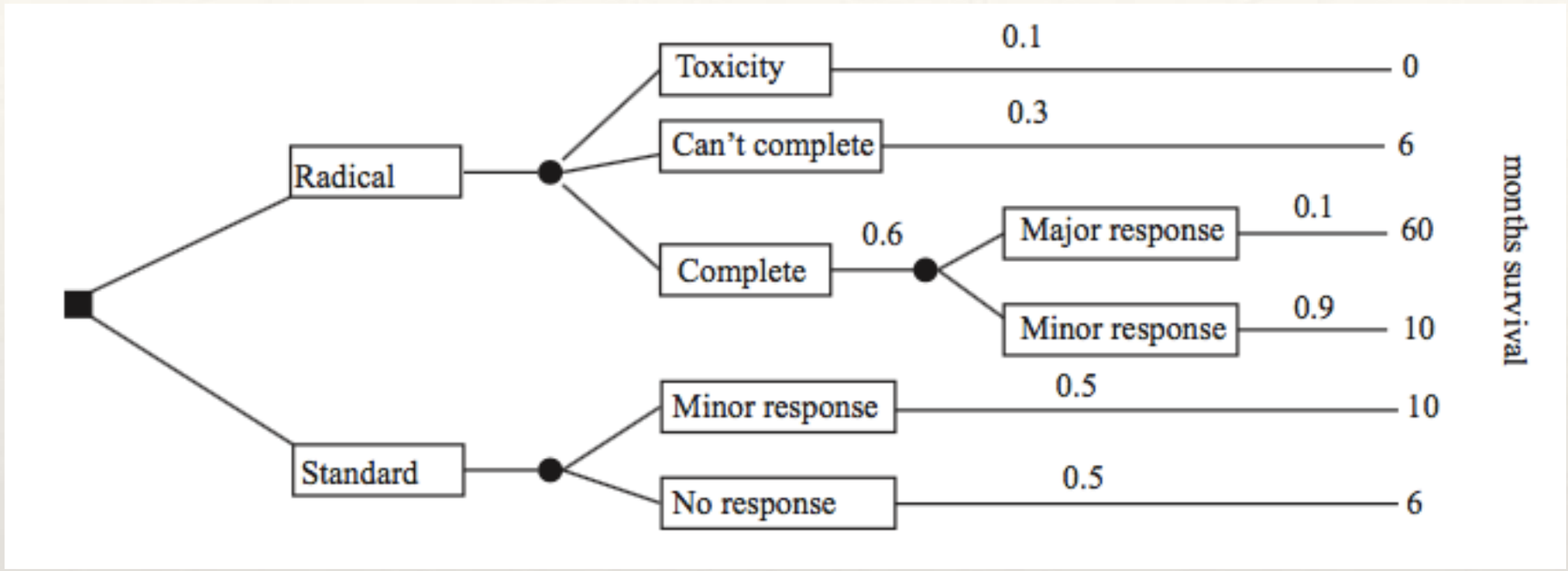
Example

Cash isn't always the quantity of interest when we are making decisions under uncertainty and using expected values as a guide. Consider the following.

Imagine you have to make a decision about how to treat a terminal disease. There are two treatment options: standard and radical. Radical treatment might kill you in and of itself with probability 0.1, or it might be so damaging that doctors stop treatment with probability 0.3 or you could complete radical treatment with probability 0.6

For any treatment there is a major, minor, or no response. The probability of major response with radical treatment is 0.1 while minor is 0.9. For standard treatment the probability of a minor response is 0.5 and no response is 0.5

Example



Survival time for radical treatment

$0.1(0) + 0.3(6) + 0.6(0.1(60) + 0.9(10)) = 10.8$ months

Survival time for standard treatment

$0.5(10)+0.5(6) = 8$ months

Useful probability distributions

Why study distributions

It's worth looking at certain broad families of random variables and their distributions because these families pop up a lot in practice. Allow us to construct models that can answer many questions

What process produced the data we see? Or what can the data we observe tell us about the underlying probabilities?

What kind of data can we expect in the future?

How should we label unlabelled data? If we have a bunch of financial transactions labelled as fraudulent and legitimate, how confidently can we label a new unlabelled transaction?

Is something we are observing an interesting effect or explainable as just random noise?

The discrete uniform distribution

If every value of a discrete random variable has the same probability, then its distribution is called a **discrete uniform distribution**

We've used this in a number of examples: heads 1, tails 0 with a fair coin. Rolling a fair die, etc.

If there are N possible values, $P(X=x) = 1/N$ if x is one of the allowable values

Bernoulli random variables

A random variable is a **Bernoulli random variable** if it takes a value of 1 with probability p and a value of 0 with probability 1-p

This is a random variable we have used for biased coin flips before

We've even derived its expectation and variance

$$E[X] = p \qquad \text{Var}[X] = p(1-p)$$

Geometric random variables

If we have a biased coin where P(H) = p and we flip this coin until heads appears, the number of flips required is a discrete random variable taking integer values greater than or equal to 1

To think of what the form of the distribution of this variable is, consider that it requires us to get n-1 tails each with probability 1-p and then one head with probability p. So we have

Distribution

$$P(\{X = n\}) = (1-p)^{n-1}p$$

p is a called **parameter** of the distribution

Expectation and variance

$$E[X] = \frac{1}{p} \qquad \text{var}[X] = \frac{1-p}{p^2}$$

In homework, geometric series, hence the name

Geometric random variables

- Not really just about coins
- It can work for any situation where we have trials characterized by success and failure
- In some cases we want to see how to model the number of successes until a failure
- Or we may flip the logic and wish to model the number of failures until a success finally occurs