Statistics 110: Introduction to Probability | Section 3, Fall 2018

Random Variables and their Distributions

TFs: Justin Zhu (justinzhu@college)

Famous Applications of Conditional Probability

Monty Hall

If we knew the location of the car, we could determine whether or not switching doors would be beneficial. This motivates us to condition on which door the car is behind. Without loss of generality, suppose we chose door 1. Let C_1, C_2, C_3 be the events that the car is behind the first, second and third doors respectively. Let W be the event that we win and receive the car. We will analyze the two situations. In both cases, we will use LOTP in the following way:

$$\begin{aligned} P(W) &= P(W|C_1)P(C_1) + P(W|C_2)P(C_2) + P(W|C_3)P(C_3) \\ &= P(W|C_1) \cdot \frac{1}{3} + P(W|C_2) \cdot \frac{1}{3} + P(W|C_3) \cdot \frac{1}{3} \end{aligned}$$

- **Strategy = Switch**
  We have that $P(W|C_1) = 0$ because if our initial guess was correct and we always switch, then we always lose. In the other 2 cases, $P(W|C_1) = P(W|C_2) = 1$ because we have chosen one of the goats, another is revealed, so by switching we are guaranteed the car. Therefore:

$$P(W) = 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3} = \boxed{\frac{2}{3}}$$

- **Strategy = Stay**
  Here, we will only win if we picked the right door to begin with, so:

$$P(W) = 1 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} = \boxed{\frac{1}{3}}$$

Gambler's Ruin

This problem is an example of *first step analysis*, where we use the Law of Total Probability, conditioned on what could happen at the first step, and look for a recursive way to express the probability of success given the first step.

In the Gambler's Ruin problem, we let p_i denote the probability of winning when you have i dollars, while your opponent has $N - i$ dollars. Using first step analysis allows us to write an equation relating the p_i's. Let W be the event that you win.

$$\begin{aligned} p_i &= P(W|\text{starting with \$}i\text{, you win round 1}) \cdot p + P(W|\text{starting with \$}i\text{, you lose round 1}) \cdot q \\ &= P(W|\text{you have \$}i+1) \cdot p + P(W|\text{you have \$ }i-1) \cdot q \\ &= p_{i+1} \cdot p + p_{i-1} \cdot q \end{aligned}$$

We also have the edge case conditions that $p_0 = 0$ and $p_N = 1$. The above is called a *difference equation*, and the closed form solution for p_i is given as follows:

$$p_i = \begin{cases} \frac{1-\left(\frac{q}{p}\right)^i}{1-\left(\frac{q}{p}\right)^N} & p \neq q \\ \frac{i}{N} & p = q = \frac{1}{2} \end{cases}$$

Simpson's Paradox

Simpson's Paradox says that X can be less successful than Y when compared within every group (Surgery Types, etc.), but still X can be more successful than Y when compared in the aggregate. This result is caused by a lurking variable, the groups, who 1) have a large impact on the success rate and 2) whose relative sizes are very different between X and Y. In statistical language, you can have that:

$$P(A|B,C) < P(A|B^c,C) \qquad \text{and} \qquad P(A|B,C^c) < P(A|B^c,C^c) \qquad \text{but} \qquad P(A|B) > P(A|B^c)$$

where A is success, B is one of the things that you are comparing (e.g. Doctors), and C is one of the groups (e.g. Surgery Types).
Here's an example:

	Doctor A	Doctor B
Applying band-aid	$^{81}/_{87}(93\%)$	$^{234}/_{270}(87\%)$
Open-heart surgery	$^{192}/_{263}(73\%)$	$^{55}/_{80}(69\%)$
Both	$^{273}/_{350}(78\%)$	$^{289}/_{350}(83\%)$

Random Variables

Formal Definition - A random variable X is a *function* mapping the sample space S into the real line.

Descriptive Definition - A random variable takes on a numerical summary of an experiment. The randomness comes from the randomness of what outcome occurs. Each outcome has a certain probability. A discrete random variable may only take on a finite (or countably infinite) number of values. Random variables are often denoted by capital letters, usually X and Y.

Distributions

A distribution describes the probability that a random variable takes on certain values. Some distributions are commonly used in statistics because they can help model real life phenomena.

PMF, CDF, and Independence

Probability Mass Function (PMF) (Discrete Only) gives the probability that a random variable takes on the value X.

$$P_X(x) = P(X = x)$$

Cumulative Distribution Function (CDF) gives the probability that a random variable takes on the value x or less

$$F_X(x_0) = P(X \leq x_0)$$

Independence - Intuitively, two random variables are independent if knowing one gives you no information about the other. X and Y are independent if for ALL values of x and y:

$$P(X = x, Y = y) = P(X = x)P(Y = y)$$

Bernoulli Distribution

Bernoulli The Bernoulli distribution is the simplest case of the Binomial distribution, where we only have one trial, or $n = 1$. Let us say that X is distributed Bern(p). We know the following:

Story. X "succeeds" (is 1) with probability p, and X "fails" (is 0) with probability $1 - p$.

Example. A fair coin flip is distributed Bern($\frac{1}{2}$).

PMF. The probability mass function of a Bernoulli is:

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

or simply

$$P(X = x) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \end{cases}$$

Binomial Distribution

Binomial Let us say that X is distributed Bin(n, p). We know the following:

Story X is the number of "successes" that we will achieve in n independent trials, where each trial can be either a success or a failure, each with the same probability p of success.

Example If Jeremy Lin makes 10 free throws and each one independently has a $\frac{3}{4}$ chance of getting in, then the number of free throws he makes is distributed Bin(10, $\frac{3}{4}$), or, letting X be the number of free throws that he makes, X is a Binomial Random Variable distributed Bin(10, $\frac{3}{4}$).

PMF The probability mass function of a Binomial is:

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

Hypergeometric

Let us say that X is distributed HGeom(w, b, n). We know the following:

Story In a population of b undesired objects and w desired objects, X is the number of "successes" we will have in a draw of n objects, without replacement.

Example 1) Let's say that we have only b Weedles (failure) and w Pikachus (success) in Viridian Forest. We encounter n of the Pokemon in the forest, and X is the number of Pikachus in our encounters. 2) The number of aces that you draw in 5 cards (without replacement). 3) You have w white balls and b black balls, and you draw b balls. X is the number of white balls you will draw in your sample.

PMF The probability mass function of a Hypergeometric is:

$$P(X = k) = \frac{\binom{w}{k}\binom{b}{n-k}}{\binom{w+b}{n}}$$

Geometric

Let us say that X is distributed Geom(p). We know the following:

Story X is the number of "failures" that we will achieve before we achieve our first success. Our successes have probability p.

Example If each pokeball we throw has a $\frac{1}{10}$ probability to catch Mew, the number of failed pokeballs will be distributed Geom$(\frac{1}{10})$.

PMF With $q = 1 - p$, the probability mass function of a Geometric is:

$$P(X = k) = q^k p$$

Discrete Distributions

Distribution	PDF and Support	Expected Value	Equivalent To
Bernoulli Bern(p)	$P(X = 1) = p$ $P(X = 0) = q$	p	Bin$(1, p)$
Binomial Bin(n, p)	$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ $k \in \{0, 1, 2, \dots n\}$	np	Sum of n independent Bern(p)
Geometric Geom(p)	$P(X = k) = q^k p$ $k \in \{0, 1, 2, \dots\}$	$\frac{q}{p}$	
Hypergeometric HGeom(w, b, n)	$P(X = k) = \binom{w}{k}\binom{b}{n-k} \Big/ \binom{w+b}{n}$ $k \in \{0, 1, 2, \dots\}$	$n\frac{w}{b+w}$	

Practice Problems

Example 1. Calvin and Hobbes.
Calvin and Hobbes play a match consisting of a series of games, where Calvin has probability p of winning each game (independently). They play with a "win by two" rule: the first player to win two games more than his opponent wins the match. Find the probability that Calvin wins the match (in terms of p) in two different ways:

(a) by conditioning, using the Law of Total Probability

(b) by interpreting the problem as a gambler's ruin problem.

Example 2. Quick Gambler's Ruin.
As in the gambler's ruin problem, two gamblers, A and B, make a series of bets, until one of the gamblers goes bankrupt. Let A start out with i dollars and B start out with $N - i$ dollars, and let p be the probability of A winning a bet, with $0 < p < \frac{1}{2}$. Each bet is for $\frac{1}{k}$ dollars, with k a positive integer, e.g., $k = 1$ is the original gambler's ruin problem and $k = 20$ means they're betting nickels. What is the probability that A wins the game?

Example 3. Mixture of Binomials.
There are two coins, one with probability p_1 of Heads and the other with probability p_2 of Heads. One of the coins is randomly chosen (with equal probabilities for the two coins). It is then flipped $n \geq 2$ times. Let X be the number of times it lands Heads.

(a) Find the PMF of X.

(b) What is the distribution of X if $p_1 = p_2$?

(c) Give an intuitive explanation of why X is not Binomial for $p_1 \neq p_2$ (its distribution is called a mixture of two Binomials). You can assume that n is large for your explanation, so that the frequentist interpretation of probability can be applied.

Example 4. Symmetry.
For the following 2 exercises, think about how symmetry may be used to avoid unnecessary calculations.

(a) Suppose X and Y are i.i.d. Bin(n, p). What is $P(X < Y)$?

(b) Can you construct two random variables X and Y both distributed Bin(3, $\frac{1}{2}$) such that $P(X = Y) = 0$?

Example 5. Counting Cards.
In the game Texas Hold'em, players combine two of their cards that are hidden to everyone else with five community cards to make the best possible five-card hand. The game is played with a standard deck of 52 cards. A flush is where all 5 cards belong to the same suit.

Suppose you are holding 2 spades in your hand, and there are 2 spades showing among the three community cards. What is the probability that you hit the flush?

Example 6. Conditional Binomial.
Suppose $X \sim \text{Bin}(n_1, p)$ and $Y \sim \text{Bin}(n_2, p)$ are independent. What is the conditional distribution of X given that $X + Y = k$, where $k \in \{0, 1, \ldots, n_1 + n_2\}$?