Chapter 2: Random Variables

Overview:

- Random variables. Discrete random variable. Probability mass function.
- Useful discrete distributions: uniform, Bernoulli, binomial, uniform, geometric, negative binomial, hypergeometric, Poisson.
- Brief calculus review (differentiation and integration).
- Continuous random variables: density, cumulative distribution function, quantile function.
- Useful continuous random variables: uniform, exponential, gamma, Weibull, normal, beta.
- Functions of a random variable.

Random variables

A random variable X gives the outcome of an experiment with a numerical sample space. Examples of random variables are:

1. Roll a 6-sided die, the outcome X is a random variable.
2. The height X of a randomly selected individual in inches is a random variable.
3. The lifetime in years X of a randomly selected smoker is a random variable. Aspects of X might be compared to Y, the lifetime of a non-smoker.
4. The number of lottery tickets X purchased until winning is a random variable.

1 and 4 are **discrete** random variables, 2 and 3 are **continuous** random variables.

Formally, a random variable is a map or function from the sample space of an experiment S to some subset of the real numbers $R \subset \mathbb{R}$.

It's probably easier to think of the random variable as having its own probably distribution with experimental outcomes taking the form $X = x$ where $x \in R$. i.e. $S = \{X = x : x \in R \subset \mathbb{R}\}$.

e.g. When rolling a 6-sided die, the possible outcomes are $X = 1$, $X = 2$, $X = 3$, $X = 4$, $X = 5$, and $X = 6$. The sample space is then $S = \{X = 1, X = 2, X = 3, X = 4, X = 5, X = 6\}$. The event that the die roll is even is given by $\{X = 2, X = 4, X = 6\}$ or more simply $X \in \{2, 4, 6\}$. We talk about the probability $P(X \in \{2, 4, 6\})$, which here equals 0.5 for a fair die.

We can denote the probability of an even roll as

- $P(\{X = 2, X = 4, X = 6\})$
- or $P(X = 2 \text{ or } X = 4 \text{ or } X = 6)$
- or $P(X \in \{2, 4, 6\})$
- or $P(X \in E)$ where $E = \{2, 4, 6\}$.

These are different ways of writing the same number ($= 0.5$ with a fair die).

e.g. The survival time in months for a pig with toxoplasmosis is given by X. The sample space is $\{X = s : s \geq 0\}$. We are interested in probabilities such as $P(X \geq 12)$, $P(X \leq 6)$, $P(X \geq 24)$.

def'n: A random variable X is **discrete** if the possible outcomes are finite or countably infinite.

- Roll a 6-sided die, let X be the result. The number of possible outcomes $R = \{1, 2, 3, 4, 5, 6\}$ is finite, so X is discrete.
- Let X be the number of lottery tickets purchased until a win. $X = 1, 2, 3, \ldots$ so X is discrete. The positive integers $R = \{1, 2, 3, \ldots\}$ are countably infinite.

def'n: The *frequency function*, or *probability mass function* (pmf) $p(x)$ of a discrete random variable X is the probability X equals x: $p(x) = P(X = x)$. The pmf completely defines a random variable X as the probability of all possible events can be computed once $p(x)$ is known.

def'n: The set of values that X can take on with positive probability is called the range of X, denoted $R = \{x \in \mathbb{R} : p(x) > 0\}$. If we know the range R, we only need to define $p(x)$ for those $x \in R$ since by definition, $p(x) = 0$ for $x \notin R$.

Here X is *random* and x is a fixed, known number.

The probability of X being in any set A is given by

$$P(X \in A) = \sum_{x:x \in A, p(x)>0} p(x) = \sum_{x:x \in A \cap R} p(x).$$

e.g. Roll a fair 6-sided die, let X be the result. Here $p(j) = 1/6$ for $j = 1, 2, 3, 4, 5, 6$. Then

$$P(X \leq 4) = p(1) + p(2) + p(3) + p(4) = \frac{2}{3}.$$

We add up probabilities of individual outcomes in the set $A = (-\infty, 4]$ that are also in $R = \{1, 2, 3, 4, 5, 6\}$.

This introduces the first random variable we'll consider, the **uniform** random variable.

def'n: A random variable X has a uniform distribution over the integers $R = \{1, 2, 3, \ldots, N\}$ if the probability function $p(j) = \frac{1}{N}$ for $j = 1, 2, 3, \ldots, N$. We denote this $X \sim U\{1, \ldots, N\}$ and say "X is distributed discrete uniform from 1 to N."

A uniform distribution is appropriate if all experimental outcomes $X \in \{1, 2, 3, \ldots, N\}$ are equally likely to happen.

e.g. N balls, numbered $1, 2, \ldots, N$, are placed in an urn and a ball randomly drawn. The number on the ball is the experimental outcome and has a uniform distribution.

Let $X \sim \text{unif}\{1, \ldots, 6\}$. What is $P(1 < X \leq 3)$?

$$P(1 < X \leq 3) = p(2) + p(3) = 2 \times \frac{1}{6} = \frac{1}{3}.$$

2.1.1 Bernoulli distribution

def'n: $X \sim \text{Bern}(p)$ if the range is $R = \{0, 1\}$ and $p(0) = 1 - p$, $p(1) = p$.

We say X has a Bernoulli distribution with parameter p. Note that for X to be properly defined that we require $0 \leq p \leq 1$.

e.g. A coin is tossed. Say if the coin is "heads" we record $X = 1$ (which happens with probability p) and for "tails" we record $X = 0$ (probability $1 - p$). Then $X \sim \text{Bern}(p)$.

The Bernoulli random variable can be used as a building block to obtain more useful random variables such as the binomial and geometric. It is also the distribution of the outcome variable, or *response*, in logistic regression.

2.1.2 Binomial distribution

Think of tossing a coin n times and recording the number of "heads." Assuming coin tosses are independent of each other (which is a good assumption), the number of heads, which can be any of $R = \{0, 1, 2, \ldots, n\}$ is said to have a binomial distribution.

Consider n independent experiments where each experiment results in a "success" with probability p and a "failure" with probability $1 - p$. The total number of successes is a binomial random variable X with parameters n and p.

def'n: $X \sim \text{bin}(n, p)$ if $R = \{0, 1, 2, \ldots, n\}$ and

$$p(x) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, 2, \ldots, n.$$

Let's work this out for $n = 4$. There are $2^4 = 16$ possible *ordered* outcomes to the experiment $\{ssss, sssf, ssfs, ssff, sfss,$ $sfsf, sffs, sfff, fsss, fssf, fsfs, fsff, ffss, ffsf, fffs, ffff\}$. The probability of, say, $\{sfsf\}$ is, using obvious notation:

$$\begin{aligned}
P(\{sfsf\}) &= P(\{s_1\} \cap \{f_2\} \cap \{s_3\} \cap \{f_4\}) \\
&= P(\{s_1\})P(\{f_2\})P(\{s_3\})P(\{f_4\}) \\
&= p(1-p)p(1-p) = p^2(1-p)^2 \\
&= p^x(1-p)^{n-x}.
\end{aligned}$$

where $x = 2$ successes, because all experiments are *mutually independent* (defined last time).

Where does the $\begin{pmatrix} n \\ x \end{pmatrix}$ come from in $p(x)$? On the previous slide we computed $P(\{sfsf\})$, the probability of two successes in a *certain, fixed order*, but the binomial random variable doesn't care about order, just the number of successes. That is,

$$\begin{aligned} P(X=2) &= P(\{ssff, sfsf, sffs, fssf, fsfs, ffss,\}) \\ &= P(\{ssff\}) + P(\{sfsf\}) + P(\{sffs\}) \\ &\quad +P(\{fssf\}) + P(\{fsfs\}) + P(\{ffss\}) \\ &= 6 \times p^2(1-p)^2 = \begin{pmatrix} 4 \\ 2 \end{pmatrix} p^2(1-p)^2. \end{aligned}$$

There are $\begin{pmatrix} 4 \\ 2 \end{pmatrix} = 6$ different possible ways to get two successes when keeping track of order. Recall that we turned this into a problem involving combinations of distinct objects by instead asking "how many combinations of size $x = 2$ are there from $\{1, 2, 3, 4\}$?" i.e. How many different slots can we place two "s" into?

In general, to derive $\begin{pmatrix} n \\ x \end{pmatrix}$ in $p(x)$ we map backwards from the range $\{0, 1, 2, \ldots, n\}$ to all possible ordered outcomes of experiments and *count* the number of orderings of x successes and $n - x$ failures.

See Figure 2.3 on page 38 for plots of the pmf $p(x)$ for $X \sim \text{bin}(10, p)$ when $p = 0.1$ or $p = 0.5$.

Example: Paraphrased from

`http://www.ninds.nih.gov/disorders/taysachs/taysachs.htm`

"Tay-Sachs is a genetic lipid storage disorder in which harmful quantities of a fatty substance called ganglioside G_{M2} build up in tissues and nerve cells in the brain. Infants with Tay-Sachs disease appear to develop normally for the first few months of life. Then, as nerve cells become distended with fatty material, a relentless deterioration of mental and physical abilities occurs. The child becomes blind, deaf, and unable to swallow. Muscles begin to atrophy and paralysis sets in. The disease occurs primarily in infants of Eastern European and Askhenazi Jewish descent."

Page 38: If a couple are both carriers of Tay-Sachs, one of their offspring has a $p = 0.25$ chance of being born with the disease. The pmf for the number of children X out of $n = 4$ that could have the disease is

$$p(x) = \left(\begin{array}{c} 4 \\ x \end{array} \right) 0.25^x 0.75^{4-x} \text{ for } x = 0, 1, 2, 3, 4.$$

See table, top of page 39 for values. Question: what is $P(X = 0)$? What is $P(X \geq 1)$? What is $P(X < 3)$?

Note $P(X \geq 1) = 1 - P(X = 0)$ because $\{0\} \cup \{1, 2, 3, 4\} = R$. That is $\{0\}$ and $\{1, 2, 3, 4\}$ are complimentary with respect to the range of possible values R, *not* all real numbers $\mathbb{R} = (-\infty, \infty)$.

2.1.3: Geometric and negative binomial distributions

Instead of counting the number of successes out of n as the binomial random variable does, the *geometric* random variable X counts the number of independent experiments *until* a success occurs, where each experiment has probability p of success.

def'n $X \sim \text{geom}(p)$ if $R = \{1, 2, 3, \ldots\}$ and $p(x) = (1-p)^{x-1}p$ for $x = 1, 2, 3, \ldots$

This one is a bit more straightforward. The sample space of possible outcomes looks like $\{s, fs, ffs, fffs, ffffs, fffffs, \ldots\}$. So, e.g.,

$$\begin{aligned} P(X=3) &= P(\{ffs\}) = P(\{f_1\} \cap \{f_2\} \cap \{s_3\}) \\ &= P(\{f_1\})P(\{f_2\})P(\{s_3\}) = (1-p)^2 p. \end{aligned}$$

See page 40 for part of the plot of a $\text{geom}\left(\frac{1}{9}\right)$ pmf.

The **negative binomial** random variable X counts the number of independent experiments until r successes occur. Note then that $R = \{r, r+1, r+2, \ldots\}$.

e.g. Let $r = 2$. The sample space of possible outcomes looks like $\{ss, fss, sfs, ffss, fsfs, sffs, \ldots\}$. Combinatorics will come into play again and your book elegantly reasons that

$$p(x) = \left(\begin{array}{c} x-1 \\ r-1 \end{array} \right) p^r (1-p)^{x-r} \text{ for } x = r, r+1, r+2, \ldots.$$

In words, in order to have r successes total, and to have one of those successes occur on the x^{th} experiment, we need to count the number of ways we can distribute the remaining $r-1$ successes among the $x-1$ slots $\{1, 2, 3, \ldots, x-1\}$.

Example: Bertrand buys a lot of generic soda. Although cheap, a can of soda is flat with probability 0.2. Consider the number of sodas X that Bertrand must drink until he's enjoyed a total of $r = 6$ that are not flat. The probability of a non-flat soda is $p = 1 - 0.2 = 0.8$.

$$p(x) = \binom{x-1}{5} 0.8^6 0.2^{x-6} \text{ for } x = 6, 7, 8, \ldots$$

What is the probability that Bertrand has enjoyed exactly $r = 6$ non-flat sodas by drinking either 6, 7, or 8 sodas?

$$\begin{aligned} P(6 \leq X \leq 8) &= p(6) + p(7) + p(8) \\ &= \binom{5}{5} 0.8^6 + \binom{6}{5} 0.8^6 0.2 + \binom{7}{5} 0.8^6 0.2^2 \\ &= \approx 0.797. \end{aligned}$$

2.1.4 Hypergeometric distribution

def'n: X is said to be hypergeometric with parameters r, n, and m, denoted $X \sim \text{hyp}(r, n, m)$ if

$$p(x) = \frac{\begin{pmatrix} r \\ x \end{pmatrix} \begin{pmatrix} n-r \\ m-x \end{pmatrix}}{\begin{pmatrix} n \\ m \end{pmatrix}} \text{ for } x = 0, 1, \ldots, m,$$

where we further define $\begin{pmatrix} s \\ t \end{pmatrix} = 0$ when $s < t$.

Place n balls in an urn of which r are white and $n-r$ are black. When choosing m balls from the urn without replacement, let X be the number of white balls. $X \sim \text{hyp}(r, n, m)$.

Note that the range of X depends on (r, n, m).

For example if $r = 2$ white balls and $n - r = 2$ black balls are placed in an urn and we draw $m = 3$ without replacement, $R = \{1, 2\}$. Why is $X = 0$ not allowed? Why is $X = 3$ or $X = 4$ not allowed? At any rate, note that defining $\begin{pmatrix} 2 \\ 3 \end{pmatrix} = \begin{pmatrix} 2 \\ 4 \end{pmatrix} = 0$ takes care of this in the definition of the pmf.

Question: what is X distributed as if balls are sampled *with* replacement? Hint: the probability of drawing a white $p = r/n$ doesn't change from draw to draw...

Where does this the pmf from? Let's look for the example $r = n - r = 2$ and a draw of $m = 3$. How many ordered outcomes have $x = 2$? $\{wwb, wbw, bww\}$. How many outcomes total? $\{wwb, wbw, bww, bbw, bwb, wbb\}$. So $P(X = 2) = 0.5$.

Think of having having the balls marked so they are distinct, e.g. W_1, W_2, B_1, B_2.

There are $\begin{pmatrix} 2 \\ 2 \end{pmatrix} = 1$ way to choose two white balls from $\{W_1, W_2\}$;
$\begin{pmatrix} 2 \\ 1 \end{pmatrix} = 2$ ways to choose one black ball from $\{B_1, B_2\}$, and
$\begin{pmatrix} 4 \\ 3 \end{pmatrix} = 4$ ways to choose $m = 3$ balls total from $\{W_1, W_2, B_1, B_2\}$.

2.1.5 Poisson distribution

def'n $X \sim \text{Pois}(\lambda)$ if

$$p(x) = \frac{\lambda^x}{x!} e^{-\lambda} \text{ for } x = 0, 1, 2, 3, \ldots$$

The Poisson distribution is used often to count the occurrence of events over time, space, or both.

- Consider the number of lip cancer cases in a year Y_i across the Scottish counties $i = 1, \ldots, 33$. Each of Y_i might be modeled as Poisson with it's own yearly rate $Y_i \sim \text{Pois}(\lambda_i)$. Poisson regression would attempt to relate lip cancer rates $\lambda_1, \ldots, \lambda_{33}$ to a covariate such as the percentage of the county population x_i involved in agricultural pursuits (i.e. that spend a lot of time outdoors).

- The number of earthquakes in a day for a fixed region in California that includes a portion of the San Andreas fault might be modeled as Poisson with rate λ where λ might be in units "earthquakes per day in the region."

- The number of armadillos killed in a day by the Ache hunters of Paraguay can be modeled with a Poisson distribution.

Note: $\sum_{x=0}^{\infty} \frac{\lambda^x e^{-\lambda}}{x!} = 1$. Why? Also, $\sum_{x=1}^{\infty}(1-p)^{x-1}p = 1$ and $\sum_{x=0}^{n} \left(\begin{array}{c} n \\ x \end{array} \right) p^x(1-p)^{n-x} = 1$.

The Poisson distribution is primarily used to count “rare” events over time or space. They are rare in the sense that if you consider a small enough interval of time, say, only one event can occur in that time period.

The Poisson distribution for X approximates the binomial distribution when $\lambda = np$ for large n and small p. The Why does the binomial distribution need to be approximated? One reason is, when n is large, computing $\begin{pmatrix} n \\ x \end{pmatrix}$ gets quite difficult, even for fast computers.

Example: At a service counter only one person can be at the service counter at a time. The probability that a customer arrives (and is served) during any given minute is 0.1. So the number of customers served in an *hour* is $X \sim \text{bin}(60, 0.1)$. We expect $np = 60 \times 0.1 = 6$ customers to be served in an hour on average, assuming that only one customer can be served in a minute.

If you hire a more speedy counter person, perhaps ten people can be serviced in a minute. Assuming people arrive at the same rate as before, the probability that one person arrives in a tenth of a minute is $0.1/10 = 0.01$. Then the number of people serviced in an hour is $X \sim \text{bin}(600, 0.01)$. Note that as before $np = 600 \times 0.01 = 6$.

We might as well give this event rate it's own Greek letter, $\lambda = 6$ people an hour. This is the service rate which stays the same; in both cases it's 6 people an hour = 0.1 person per minute = 0.01 person per tenth of a minute.

In the limit, as $n \to \infty$ and $p \to 0$, but holding $\lambda = np$ constant,

$$\binom{n}{x} p^x (1-p)^{n-x} \longrightarrow \frac{e^{-np}(np)^x}{x!} = \frac{e^{-\lambda}\lambda^x}{x!},$$

and the Poisson distribution approximates the binomial. The plots I showed you gives the pmfs for two binomials and a Poisson for a fixed $\lambda = np$. The pmfs are almost indistinguishable. See also Example A on page 44.

Example: A researcher, Dr. Johnson, posits the hypothesis that there exist two kinds of people: those that mosquitoes like, A, and those that mosquitoes do not like A^C. Dr. Johnson thinks that the two types of people occur with equal frequencies, i.e. the probability that Blanche is favored by mosquitoes is $P(A) = 0.5$.

Dr. Johnson has devised an experiment in which he places a person in an enclosed plastic booth filled with hundreds of mosquitoes for one minute. From empirical evidence, Dr. Johnson has concluded that the number of mosquito bites X a type A person receives is $X \sim \text{Pois}(11.3)$ bites in a minute and those a type A^C person receives is $X \sim \text{Pois}(2.7)$.

Say Blanche tries Dr. Johnson's experiment and receives $X = 8$ bites. Is Blanche more likely to be type A or type A^C?

Answer: Using Bayes' rule we have

$$
\begin{aligned}
P(A|X=8) &= \frac{P(X=8|A)P(A)}{P(X=8|A)P(A)+P(X=8|A^C)P(A^C)} \\
&= \frac{\frac{11.3^8 e^{-11.3}}{8!}0.5}{\frac{11.3^8 e^{-11.3}}{8!}0.5+\frac{2.7^8 e^{-2.7}}{8!}0.5} \approx 0.945.
\end{aligned}
$$

What kind of distribution best models the following scenarios?

- We are interested in the number of children born until malformation of the jaw is observed in an isolated village where refined sugar was only recently introduced.
- A carton of 12 eggs is purchased. We are interested in the number cracked out of 12.
- The number of earthquakes in a 24 hour period in a certain seismically active part of California is counted.
- On a sprawling ranch with 327 Hereford and 480 Guernsey, 20 cattle are randomly rounded up to be checked for Brucellosis. We are interested in the number of Hereford out of the 20.
- A fair die is rolled; X is the outcome.

Homework, Chapter 2: 11, 12*, 13, 14*, 15, 16 (just think about it), 20 (think about it), 21 (think about it), 22*, 25, 26*, 27

Coming up: continuous random variables, densities, cumulative distribution functions, calculus review, functions of random variables.