Discrete Probability Distributions

STA 198: Introduction to Health Data Science

Yue Jiang

June 01, 2020

The following material was used by Yue Jiang during a live lecture.

Without the accompanying oral comments, the text is incomplete as a record of the presentation.

What is a random variable?

A random variable is a quantity whose value depends on the outcome of a random event

- Traditionally, we use capital letters X, Y, Z to denote random variables
- The values that random variables take are in lowercase x, y, z
- So, the probability that random variable X has the value x is denoted by $P(X = x)$

Random variables encountered in this class will be either discrete or continuous

Discrete random variables

Discrete random variables are those that can take on a countable number of potential values (could be countably infinite!), each associated with a probability of occurring

This list of possible values and probabilities is the probability distribution* for the discrete random variable in question

Probability distributions let us investigate how likely events may be

* For discrete random variables, you may also see the term *probability mass function*, which is the same thing

Probability distribution of births in the US


NCHS Data Brief ■ No. 346 ■ July 2019

Births in the United States, 2018

Joyce A. Martin, M.P.H., Brady E. Hamilton, Ph.D., and Michelle J.K. Osterman, M.H.S.

Let X be the random variable for birth status in the US. Its probability distribution may be given by

Event	Probability
$X = pre$	0.10
$X = early$	0.27
$X = full$	0.57
$X = late/post$	0.06

Probability distribution of births in the US

Event	Probability
$X = pre$	0.10
$X = early$	0.27
$X = full$	0.57
$X = late/post$	0.06

There are three rules for discrete probability distributions:

- Outcomes must be disjoint
- The probability of each outcome must be ≥ 0 and ≤ 1
- The sum of the outcome probabilities must add up to 1

Bernoulli random variables

Some "types" of random variables come up very often. Consider a dichotomous (two-level) random variable X:

- Dead or alive
- Current smoker or not

This is known as a Bernoulli random variable, and has a probability of "success" denoted by p

Bernoulli random variables

With the notation that event $X = 1$ is a "success" and $X = 0$ is a "failure," $P(X = 1) = p$ and $P(X = 0) = 1 - p$

- Fair coin flip: Let the event $X = 1$ denote heads.

Event	$P(\text{Event})$
$X = 1$	0.5
$X = 0$	0.5

- 2018 US births: Let the event $X = 1$ denote preterm birth.

Event	$P(\text{Event})$
$X = 1$	0.1
$X = 0$	0.9

A "success" is not necessarily positive – if we are interested in the probability of dying, a "success" would be death

Extending the Bernoulli distribution

Suppose we randomly select two independent US births in 2018, and Z is a new random variable that represents the number of preterm births among them

Z can be 0, 1, or 2:

First Birth X	Second Birth X	Number of Preterm Births Z	Probability of These Outcomes
0	0	0	
1	0	1	
0	1	1	
1	1	2	

Extending the Bernoulli distribution

Let X_1 be 1 if the first birth is preterm, and X_2 be 1 if the second birth is preterm, and 0 otherwise.

Because these two births are independent, then

$$\begin{aligned} P(\text{both preterm}) &= P(X_1 = 1 \cap X_2 = 1) \\ &= P(X_1 = 1)P(X_2 = 1|X_1 = 1) \\ &= P(X_1 = 1)P(X_2 = 1) \\ &= p * p = 0.1 \times 0.1 = 0.01. \end{aligned}$$

Extending the Bernoulli distribution

Back to the table, row 1 is given by
$P(X_1 = 0 \cap X_2 = 0) = P(X_1 = 0)P(X_2 = 0) = (1 - p)(1 - p)$

First Birth X_1	Second Birth X_2	Number of Preterm Births Z	Probability of These Outcomes
0	0	0	$(1 - p)(1 - p) = 0.81$
1	0	1	
0	1	1	
1	1	2	

Extending the Bernoulli distribution

Row 2 is given by

$$P(X_1 = 1 \cap X_2 = 0) = P(X_1 = 1)P(X_2 = 0) = p(1-p)$$

First Birth X_1	Second Birth X_2	Number of Preterm Births Z	Probability of These Outcomes
0	0	0	$(1-p)(1-p) = 0.81$
1	0	1	$p(1-p) = 0.09$
0	1	1	
1	1	2	

Extending the Bernoulli distribution

Row 3 is given by
$P(X_1 = 0 \cap X_2 = 1) = P(X_1 = 0)P(X_2 = 1) = (1 - p)(p)$

First Birth X_1	Second Birth X_2	Number of Preterm Births Z	Probability of These Outcomes
0	0	0	$(1 - p)(1 - p) = 0.81$
1	0	1	$p(1 - p) = 0.09$
0	1	1	$(1 - p)p = 0.09$
1	1	2	

Extending the Bernoulli distribution

Finally, row 4 is given by
$P(X_1 = 0 \cap X_2 = 0) = P(X_1 = 0)P(X_2 = 0) = p \times p$

First Birth X_1	Second Birth X_2	Number of Preterm Births Z	Probability of These Outcomes
0	0	0	$(1-p)(1-p) = 0.81$
1	0	1	$p(1-p) = 0.09$
0	1	1	$(1-p)p = 0.09$
1	1	2	$p \times p = 0.01$

Extending the Bernoulli distribution

Thus, the probability distribution of number of preterm births out of two independent births is given by

	$z = 0$	$z = 1$	$z = 2$
$P(Z = z)$	0.81	0.18	0.01

Extending the Bernoulli distribution

Now let Z be the random variable corresponding to the number of preterm births among 3 independently sampled births

First Birth X_1	Second Birth X_2	Third Birth X_3	Number of Preterm Births Z	Probability
0	0	0	0	
1	0	0	1	
0	1	0	1	
0	0	1	1	
1	1	0	2	
1	0	1	2	
0	1	1	2	
1	1	1	3	

Extending the Bernoulli distribution

Row 1 is given by

$$P(X_1 = 0 \cap X_2 = 0 \cap X_3 = 0) = (1-p)(1-p)(1-p)$$

First Birth X_1	Second Birth X_2	Third Birth X_3	Number of Preterm Births Z	Probability
0	0	0	0	0.729
1	0	0	1	
0	1	0	1	
0	0	1	1	
1	1	0	2	
1	0	1	2	
0	1	1	2	
1	1	1	3	

Extending the Bernoulli distribution

Row 2 is given by

$P(X_1 = 1 \cap X_2 = 0 \cap X_3 = 0) = p(1-p)(1-p)$

First Birth X_1	Second Birth X_2	Third Birth X_3	Number of Preterm Births Z	Probability
0	0	0	0	0.729
1	0	0	1	0.081
0	1	0	1	
0	0	1	1	
1	1	0	2	
1	0	1	2	
0	1	1	2	
1	1	1	3	

Extending the Bernoulli distribution

...et cetera

First Birth X_1	Second Birth X_2	Third Birth X_3	Number of Preterm Births Z	Probability
0	0	0	0	0.729
1	0	0	1	0.081
0	1	0	1	0.081
0	0	1	1	0.081
1	1	0	2	0.009
1	0	1	2	0.009
0	1	1	2	0.009
1	1	1	3	0.001

Extending the Bernoulli distribution

If we randomly sample 3 births, what is the chance 2 are preterm?
$0.009 + 0.009 + 0.009 = 0.027$ (why?)

First Birth X_1	Second Birth X_2	Third Birth X_3	Number of Preterm Births Z	Probability
0	0	0	0	0.729
1	0	0	1	0.081
0	1	0	1	0.081
0	0	1	1	0.081
1	1	0	2	0.009
1	0	1	2	0.009
0	1	1	2	0.009
1	1	1	3	0.001

Extending the Bernoulli distribution

Thus, the probability distribution of number of preterm births out of three independent births is given by

	$z = 0$	$z = 1$	$z = 2$	$z = 3$
$P(Z = z)$	0.729	0.243	0.027	0.001

(Verify: are these disjoint events, with probabilities in [0, 1], that all sum to 1?)

Extending the Bernoulli distribution

If we randomly sample 4 births, what is the probability distribution for the number of preterm births?

To build a similar table would start to become egregious. Luckily, there is a formula for this probability distribution

The binomial distribution

The binomial distribution gives the probability of k "successes" from a sequence of n independent Bernoulli trials. There are three assumptions:

1. There is a fixed number of trials n, each of which is a Bernoulli random variable
2. The outcomes of the n trials are independent
3. The probability of success, p, is the same for each of these trials

The binomial distribution

If X has a binomial distribution, then

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

The binomial distribution

If we randomly sample 3 births, what is the chance 2 are preterm?

1. There is a fixed number of trials, each of which is a Bernoulli random variable
2. The outcomes of the trials are independent
3. The probability of success is the same for each of these trials

Thus, $Z \sim Binom(3, 0.1)$.

$$P(Z = 2) = \binom{3}{2} 0.1^2 (1 - 0.1)^{3-2}$$

$$= \frac{3!}{2!(3-2)!} \times 0.1^2 \times 0.9^1 = 0.027$$

The Poisson distribution

- Discrete distribution taking on possible values 0, 1, 2, 3, ..., ∞
- Often used to model counts or rare events
- Much like the binomial distribution, requires a few assumptions

The Poisson distribution

The Poisson distribution gives the probability that k events occur in a given "interval." There are four assumptions:

1. Within any interval, k may take on values 0, 1, 2, 3, ..., ∞
2. Each event occurs independently, both within the same interval, and between intervals
3. The average rate at which events occur in an interval, λ, is constant
4. Two events cannot occur simultaneously

What is an "interval"?

The Poisson distribution

If X has a Poisson distribution, then

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

The Poisson distribution

Suppose on average, there are 1.5 deaths due to Alzheimer's Disease in a town each year. For a one-year period in this town, what is the chance that two or more people die from Alzheimer's?

1. Within any interval, the number of AD deaths can range from 0 to ∞ (*technically* not true, but close enough)
2. One individual dying of Alzheimer's does not affect the chance of another person dying of Alzheimer's
3. The AD death rate is constant in this town
4. Two AD deaths cannot occur at the same time (we can always subdivide time intervals such that only one person experiences this event in a given sub-interval)

The Poisson distribution

Thus, $Z \sim Pois(1.5)$, and $P(Z = 0)$, $P(Z = 1)$, and $P(Z \geq 2)$ are disjoint events. So, $P(Z \geq 2) = 1$ - $(P(Z = 0) + P(Z = 1))$, where

$$P(Z = 0) = \frac{1.5^0 \times e^{-1.5}}{0!} \approx 0.223$$

$$P(Z = 1) = \frac{1.5^1 \times e^{-1.5}}{1!} \approx 0.335$$

And so $P(Z \geq 2) \approx 1 - 0.223 - 0.335 = 0.442$.

What about other interval lengths?

Suppose we have a count random variable that follows a Poisson distribution.

Since each event is independent of others and the rate λ is constant, **the probability that an event occurs within an interval is proportional to the length of that interval**.

E.g., we would expect twice the number of events to occur in an interval of twice the length; we would expect 1/9 times the number of events to occur in an interval 9 times as small; and so on.

Participation What about other interval lengths?

Suppose on average, there are 1.5 deaths due to Alzheimer's Disease in a town each year.

1. What is the average one-**month** rate of deaths due to Alzheimer's disease in this town?
2. For any given one-*month* period in this town, what is the probability that exactly one person dies from Alzheimer's? (Just the expression is fine)

For $X \sim Pois(\lambda)$,

$$P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Expectation and variance

Now that we've defined random variables and explored a few distributions, we might be interested in some other aspects of their distributions.

Suppose we are interested in the probability distribution corresponding to the number of preterm births in a random sample of five independent US births

- How many preterm births should we expect?
- How "spread out" would this distribution be?

Expected value

The expected value of a discrete random variable X is a weighted average of the possible outcomes:

$$E(X) = \sum_{\text{all } x} x \cdot P(X = x)$$

This is a property of the *distribution* of X, not of the random variable itself

Properties of expected values

Let X and Y be random variables and c be a numeric constant

- $E(c) = c$
- $E(X + c) = E(X) + c$
- $E(cX) = c \cdot E(X)$
- $E(X + Y) = E(X) + E(Y)$
- $E(X - Y) = E(X) - E(Y)$
- If X and Y are independent, $E(XY) = E(X)E(Y)$

Expected value of a function of X

We can also find the expected value of a function of the random variable:

$$E(f(X)) = \sum_x f(x) \cdot P(X = x)$$

It is an average of the values, weighted by their probability of occurrence

Variance

The variance of X tells us how close values tend to be to its expectation:

$$Var(X) = E\left((X - E(X))^2\right)$$

By the last slide, for $f(X) = (X - E(X))^2$:

$$Var(X) = \sum_{\text{all } x} (x - E(X))^2 \cdot P(X = x)$$

Thus, the variance is the expected squared deviation of a random variable from its expectation

Properties of variances

Let X and Y be random variables and c be a numeric constant

- $Var(c) = 0$
- $Var(X + c) = Var(X)$
- $Var(cX) = c^2 \cdot Var(X)$
- If X and Y are independent, $Var(X + Y) = Var(X) + Var(Y)$
- If X and Y are independent, $Var(X - Y) = Var(X) + Var(Y)$

Properties of Bernoulli and binomial random variables

For $Z \sim Bern(p)$, $E(Z) = p$ and $Var(Z) = p(1-p)$

For $Z \sim Binom(n, p)$, $E(Z) = np$ and $Var(Z) = np(1-p)$

For $Z \sim Pois(\lambda)$, $E(Z) = Var(Z) = \lambda$