Random Variables and Discrete Probability Models

Ryan Miller

Outline

1. Defining a *random variable*
2. Discrete probability models
3. Expected value and standard deviation

Random Variables

- We've been studying *probability* to understand the possible *outcomes* of a *random process*
  - Two important random processes are *sampling from a population*, and *assigning treatment/control groups*

Random Variables

- We've been studying *probability* to understand the possible *outcomes* of a *random process*
  - Two important random processes are *sampling from a population*, and *assigning treatment/control groups*
- Statisticians use a **random variable** to represent the *unknown numeric outcome* of a random process
  - Like any variable, you can think of a random variable, such as X, as a written placeholder for an unknown numerical value

Random Variables

- Consider the random process of flipping a fair coin
  - Because random variables must involve a numeric outcome, we can use the value "1" to represent the outcome "heads" and "0" to represent the outcome "tails"

Random Variables

- Consider the random process of flipping a fair coin
  - Because random variables must involve a numeric outcome, we can use the value "1" to represent the outcome "heads" and "0" to represent the outcome "tails"
  - We could've also mapped tails to 1 and heads to 0 without any consequence (so long as we keep track of what is what)

Random Variables

- Consider the random process of flipping a fair coin
  - Because random variables must involve a numeric outcome, we can use the value "1" to represent the outcome "heads" and "0" to represent the outcome "tails"
  - We could've also mapped tails to 1 and heads to 0 without any consequence (so long as we keep track of what is what)
- We can now define X as a random variable
  - $X = 1$ if "heads" is observed, and $X = 0$ if "tails" is observed

Probability Models (example)

- After each touchdown in the National Football League (NFL), the scoring team gets to choose between a 1-pt and 2-pt attempt to earn additional points (on top of the 6 given for scoring a touchdown)

Probability Models (example)

- After each touchdown in the National Football League (NFL), the scoring team gets to choose between a 1-pt and 2-pt attempt to earn additional points (on top of the 6 given for scoring a touchdown)
- We can use a random variable X to denote the number of total points the team earns from a touchdown
  - Recognize X represents a numeric outcome that is *unknowable in advance*

Probability Models (example)

- After each touchdown in the National Football League (NFL), the scoring team gets to choose between a 1-pt and 2-pt attempt to earn additional points (on top of the 6 given for scoring a touchdown)
- We can use a random variable X to denote the number of total points the team earns from a touchdown
  - Recognize X represents a numeric outcome that is *unknowable in advance*
- Since a rule change in 2015, 9.6% of touchdowns were accompanied by zero additional points, 86.5% resulted in one additional point, and 3.9% resulted in two additional points

Probability Models (example)

- After each touchdown in the National Football League (NFL), the scoring team gets to choose between a 1-pt and 2-pt attempt to earn additional points (on top of the 6 given for scoring a touchdown)
- We can use a random variable X to denote the number of total points the team earns from a touchdown
  - Recognize X represents a numeric outcome that is *unknowable in advance*
- Since a rule change in 2015, 9.6% of touchdowns were accompanied by zero additional points, 86.5% resulted in one additional point, and 3.9% resulted in two additional points
  - Based upon these data, we might consider following **probability model** for X:

X	6	7	8
$P(X = x)$	0.096	0.865	0.039

Practice

- In Lab #3 you considered sampling from a sock warehouse that contains hundreds of thousands of socks that were 20% are blue, 50% are grey, and 30% are black.
- Letting the random variable X denote the number of black socks in sample of size $n = 2$, find a probability model for X (Hint: write out the sample space and find a probability for each member, and remember the complement rule)

Practice (solution)

Probability model for the number of black socks in a sample of size $n = 2$:

X	0	1	2
P(X = x)	0.49	0.42	0.09

Using a Probability Model

Probability models are useful because they help us understand a few key aspects of a random process:

1) **Expected Value**, or the "average" numeric outcome
2) **Variance**, or the total amount that the numeric outcomes vary from their *expected value*
3) **Standard Deviation**, or the "average" amount that numeric outcomes vary from their *expected value*

Expected Value

- The **expected value** of a random variable is denoted $E(X)$
- It describes the *expected result*, which is the sum of each possible outcome weighted by its probability

X	6	7	8
P(X = x)	0.096	0.865	0.039

- For a randomly chosen NFL touchdown, $E(X) = 6 * 0.096 + 7 * 0.865 + 8 * 0.039 = 6.94$ points

Practice

For the sock factory example, find the *expected number* of black socks in a sample of size $n = 2$.

X	0	1	2
P(X = x)	0.49	0.42	0.09

Practice (solution)

- $E(X) = 0.49 * 0 + 0.42 * 1 + 0.09 * 2 = 0.6$
- Notice how this happens be $n * P(\text{black}) = 2 * 0.3$, which is not a coincidence

Variance

Returning to the NFL example, to understand how each possible outcome (6, 7, or 8 pts) varies from the expected outcome (6.94 pts) we can calculate their *squared deviations*

Points	6	7	8
Deviation	(6-6.94)^2	(7-6.94)^2	(8-6.94)^2

If we add these squared deviations, weighted by their probabilities, we get **variance**:

$$\text{Var}(X) = 0.096*(6-6.94)^2+0.865*(7-6.94)^2+0.039*(8-6.94)^2 = 0.13$$

Variance

Returning to the NFL example, to understand how each possible outcome (6, 7, or 8 pts) varies from the expected outcome (6.94 pts) we can calculate their *squared deviations*

Points	6	7	8
Deviation	(6-6.94)^2	(7-6.94)^2	(8-6.94)^2

If we add these squared deviations, weighted by their probabilities, we get **variance**:

$$Var(X) = 0.096*(6-6.94)^2+0.865*(7-6.94)^2+0.039*(8-6.94)^2 = 0.13$$

Standard Deviation

Taking the square-root of the variance, we have the **standard deviation**, or the *average deviation* of outcomes from the expected value:

$$\text{SD}(X) = \sqrt{\text{Var}(X)} = \sqrt{0.13} = 0.36$$

So, we expect the average deviation (from the expected value of 6.94) of a touchdown to be 0.36 pts (not much variation)

Practice

For the sock factory example, find the *standard deviation* for the number of black socks in a sample of size $n = 2$.

X	0	1	2
P(X = x)	0.49	0.42	0.09

Practice (solution)

- The variance of X is:
  $0.49 * (0 - 0.6)^2 + 0.42 * (1 - 0.6)^2 + 0.09 * (1 - 0.6)^2 = 0.258$
- The standard deviation of X is: $\sqrt{0.258} = 0.508$
- This supports the notion that a samples containing either 0 or 1 black socks are likely

The Binomial distribution

- The examples we've considered have been small enough to allow us to write out the probability model using a table
  - For more complex examples, it's common to use a *mathematical function* to model these probabilities

The Binomial distribution

- The examples we've considered have been small enough to allow us to write out the probability model using a table
  - For more complex examples, it's common to use a *mathematical function* to model these probabilities
- For the sampling of *binary categorical outcomes* (such as the whether or not a sock is black), the **binomial distribution** can be used:

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

Here, $\binom{n}{x}$ describes the number of way's x "successes" can occur in a sample of size n, and p describes the probability of a "success" for each case

Comments (binomial distribution)

- In this class, I will not ask you to work directly with the binomial distribution because we can *approximate* it with a simpler probability model (the Normal model!)
- Nevertheless, it's an important model to be aware of in case you encounter and "exact binomial test" in the future