Introduction to Random Variables

Gregory M. Shinault

Goal for this Lecture

Learn some of the basic concepts surrounding random variables. This will include the following topics.

1. The definition of a random variable, and its distribution
2. Discrete random variables and their probability mass functions

The material of this lecture roughly corresponds to Section 1.5 of the textbook.

Random Variable: It is Just a Function

Definition: A *random variable* (RV) is a function whose domain is a sample space and codomain is the set of real numbers, $X : \Omega \to \mathbb{R}$.

Why RVs?

1. We cannot do conventional algebra on sample spaces.
2. We cannot always identify the sample space.

Random Variables, Explicit Sample Space

We roll 2 dice. Let X be their sum.

0. Give the sample space for the random experiment, and describe the random variable in terms of elements from the sample space.
1. $X = 7$ is an event. Write out this event as a set and find its probability.
2. Identify all the values X can take and compute the probability that X takes each value.
3. Find $\mathbb{P}(X \leq 4)$.

The Distribution of a Random Variable

The **(probability) distribution** of a random variable X is the collection of probabilities $\mathbb{P}(X \in B)$ for any set B of real numbers.

There are many possible ways to define the distribution of a RV. We will use the word "distribution" to refer to all of them.

Discrete Random Variables

Vocabulary

For now we only look at random variables with a countable range.

Definition: A random variable X is called *discrete* if the range of X is countably infinite or finite.

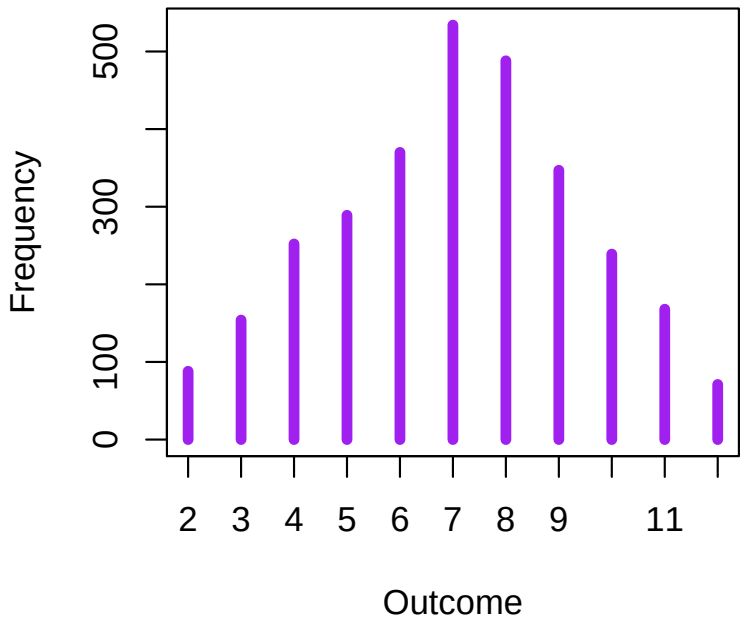
Definition: For each value $k \in \text{Range}(X)$, we define the *probability mass function* (PMF) of X as

$$p_X(k) = \mathbb{P}(X = k).$$

Comment: The PMF is analogous to the relative frequency bar chart for datasets.

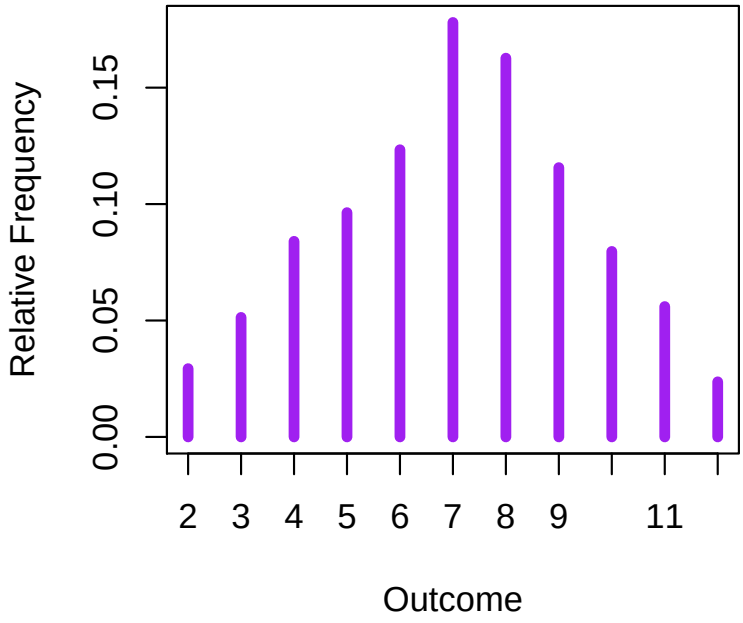
PMF vs. Relative Frequency Bar Chart

```
## diceSum
##   2   3   4   5   6   7   8   9  10  11  12
##  88 154 252 289 370 534 488 347 239 168  71
```



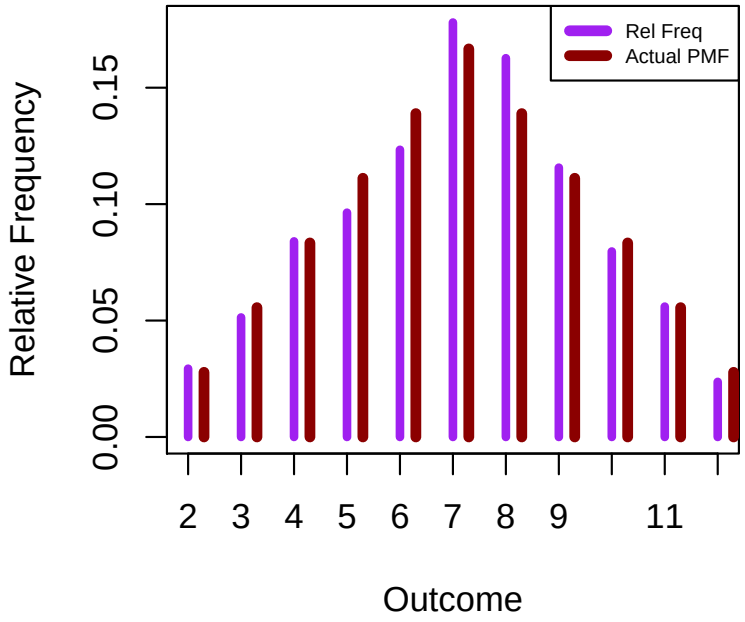
PMF vs. Relative Frequency Bar Chart

```
## diceSum
##   2   3   4   5   6   7   8   9  10  11  12
##  88 154 252 289 370 534 488 347 239 168  71
```



PMF vs. Relative Frequency Bar Chart

```
## diceSum
##   2   3   4   5   6   7   8   9  10  11  12
##  88 154 252 289 370 534 488 347 239 168  71
```



PMF vs. Relative Frequency Bar Chart

The Takeaway Lesson: The PMF tells us how frequently an outcome should occur in repeated simulations of the random variable.

Random Variables, Implicit Sample Space

Suppose the average number of customers that come to the antique shop I own during the lunch hour is 4. We can model this with X as a random variable with probability mass function

$$\mathbb{P}(X = k) = p_X(k) = e^{-4}\frac{4^k}{k!} \text{ for } k = 0, 1, 2, \ldots$$

1. First verify that this is a valid pmf. That is, check that for all k that $p_X(k) \geq 0$ and the probability for all outcomes of X sum to 1.
2. Compute the probability that I have more than 2 customers during the lunch hour.

A Mixed Discrete/Continuous RV

We throw a dart at a circular board that has radius 12 inches. There is a ring that is 0.5 inches thick, with the center halfway from the center of the board to the edge in which the point value of your throw is worth twice as much. Let X be the distance from where your dart lands to this ring. Compute $\mathbb{P}(X \leq 2)$.

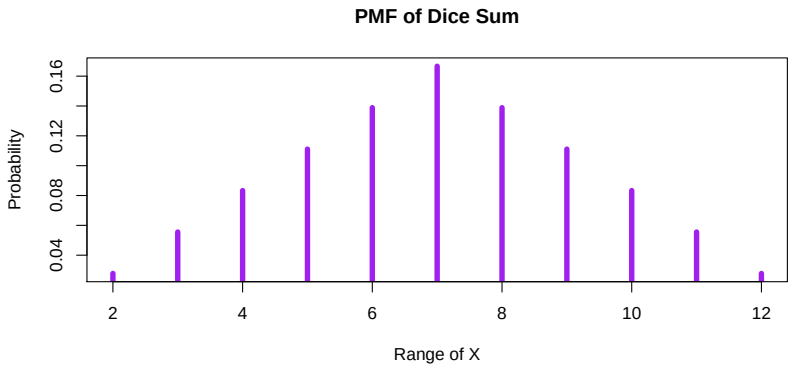
Comment: X is an example of a random variable that is a mixture of discrete and continuous. $\mathbb{P}(X = 0) > 0$, but for all other values $\mathbb{P}(X = d) = 0$.

Visualization for RVs

- To understand conventional functions $f : \mathbb{R} \to \mathbb{R}$ we usually graph them.
- The domain of RVs (sample spaces) do not have a consistent structure.
- Thus, no graphs.
- Next best thing: graph the PMF.

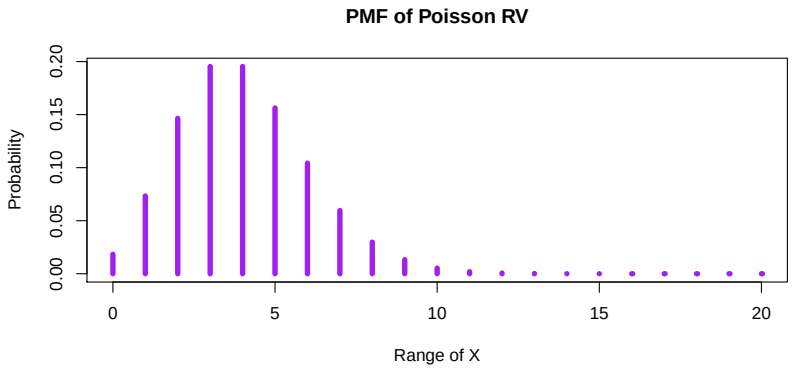
PMF graph

```
ranX <- 2:12
pmfX <- (1/36) * c(1, 2, 3, 4, 5, 6, 5, 4, 3,
    2, 1)
plot(ranX, pmfX, type = "h", col = "purple", lwd = 5,
    xlab = "Range of X", ylab = "Probability",
    main = "PMF of Dice Sum")
```



PMF graph

```
ranX <- 0:20
pmfX <- dpois(ranX, lambda = 4)
plot(ranX, pmfX, type = "h", col = "purple", lwd = 5,
    xlab = "Range of X", ylab = "Probability",
    main = "PMF of Poisson RV")
```



Applied Example

Tweet Lengths

- Let X be the length of a randomly selected tweet. What is it's PMF?
- The issue: There is no clear theoretical/mathematical approach.
- Idea: Collect a lot of tweets, count the length of each tweet, create a frequency table of tweet length.
- I did this with some python scripts. One to collect raw tweets, one to process the tweets.
- I wrote the results to a file called "tweet_length.csv".

Frequency Table

We can use R to look at this directly if we like.

```
tweet_data <- read.csv("tweet_length.csv", header = FALSE,
    col.names = c("TweetLength", "Freq"), colClasses = c("integer",
        "numeric"))
head(tweet_data)
```

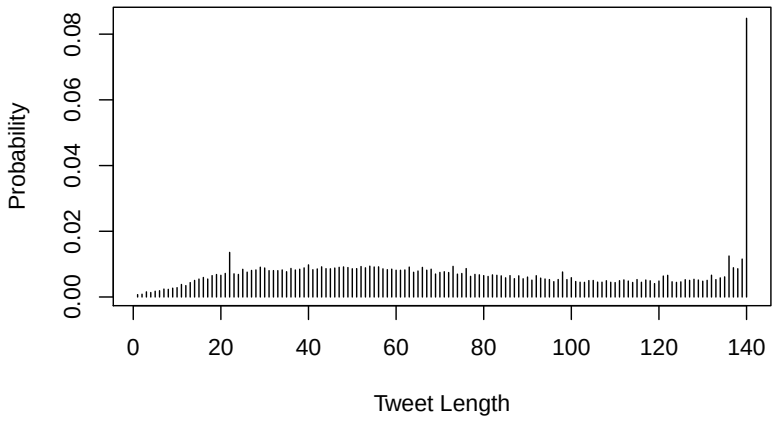
```
##   TweetLength         Freq
```

```
## 1           1 0.0007181124
## 2           2 0.0008434971
## 3           3 0.0015616095
## 4           4 0.0013564345
## 5           5 0.0017439872
## 6           6 0.0019035678
```

Frequency Bar Chart

Just looking at the PMF as a table is not helpful. We should graph it.

```
plot(1:140, tweet_data$Freq, type = "h", xlab = "Tweet Length",
    ylab = "Probability")
```

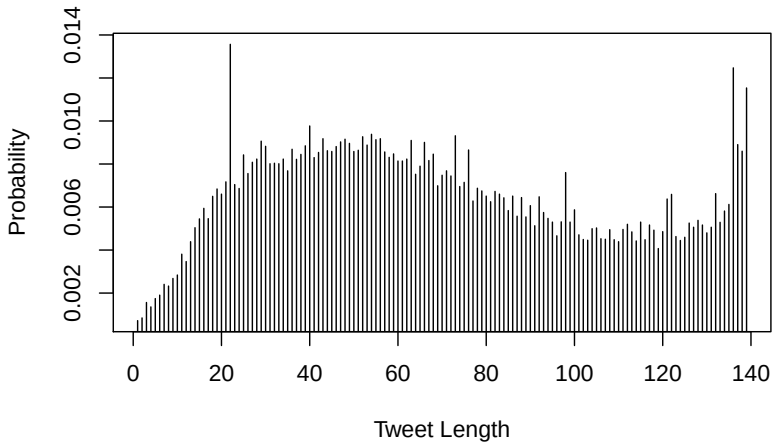


Definitive peak at $k = 140$, probably due to people running out of space.

Trimming Outliers

We can ignore $k = 140$ for additional perspective.

```
plot(1:139, tweet_data$Freq[1:139], type = "h",
    xlab = "Tweet Length", ylab = "Probability")
```



Why is $k = 22$ so big? That seems strange.

Investigation

- Write a script to print tweets of length 22 from the collection of raw tweets.
- Many are of the form "`http://t.co/ABCDEFGHIJ`".
- If someone just posts a link, it is length 22 by default.
- Conclusion: People post a lot of links on twitter.
- Larger punchline: The PMF is not any sort of deep theory, but inspecting the closely related bar chart can reveal some interesting insights in practice.

The Wrap Up

Summary

1. A random variable is just a function with a sample space for a domain
2. The PMF is the most important tool for computing probabilities for discrete RVs
3. To visualize outcomes for a discrete RV, graph the PMF