Module C: Understanding the Data-Generating Process

Slide Deck C1:

Discrete Random Variables

The section in which we learn start to focus on the random variables we measure, and the data-generating process underlying them. Here, we introduce distributions that take on specific possible values, discrete random variables. Beyond that, we cover population parameters and begin thinking about how we can use the sample to estimate them.

Today's Objectives

By the end of this slidedeck, you should

1. explain what a random variable is
2. understand the difference between discrete and continuous (random) variables
3. know the purpose of the probability mass function (pmf)
4. explain the three requirements for a function to be a pmf
5. calculate probabilities using the pmf
6. determine the sample space of a distribution
7. calculate the expected value and variance of a distribution

Random Variables

Definition

A **random variable** is a variable whose numeric value is determined by the outcome of a probability experiment.

Examples

- a statistician's favorite ice cream flavor
- a student's level of approval of a Congressional decision
- the year a Knox College professor is born
- the number of pages read by a student each night

Note: Random variables have (or follow) probability distributions. This fact allows us to *understand* the randomness of a random variable… **and of our sample**.

Properties of Probability Distributions

There are three requirements for a function to be a probability mass function:

1. All of the probabilities are between 0 and 1, inclusive.

$$0 \leq \mathbb{P}[\ X = x\ ] \leq 1$$

2. The sum of the probabilities is 1.

$$\sum_{x \in S} \mathbb{P}[\ X = x\ ] = 1$$

3. The probability of a union is no more than the sum of the individual probabilities.

$$\mathbb{P}[\ A \cup B\ ] \leq \mathbb{P}[\ A\ ] + \mathbb{P}[\ B\ ]$$

Quick Coin Example

Let us create a probability mass function for this experiment:

Flip a coin three times and count the number of heads flipped.

Solution
The first step is to determine the possible outcomes, which is called the "**sample space**."

- From the description of the experiment, these are the only outcomes possible:

$$S = \{0, 1, 2, 3\}$$

Quick Coin Example

Let us create a probability mass function for this experiment:

Flip a coin three times and count the number of heads flipped.

Solution. . .
The second step is to determine the probability of each of the elements of the sample space.

- To do this, we will rely on two assumptions:
  - the coin is fair
  - the flips are independent.

Quick Coin Example

Let us create a probability mass function for this experiment:

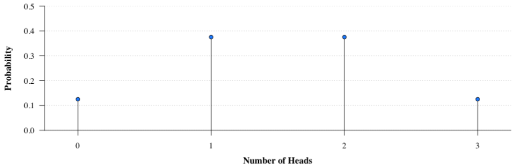
Flip a coin three times and count the number of heads flipped.

Solution. . .
If these are true, then here are the possible outcomes of three flips. A **table** is just one way of showing the probability mass function.

Heads	Flip Outcomes	Probability	
0	TTT	$\mathbb{P}[\ X = 0\ ] = 1 \times 1/8 = 1/8$	$= 0.125$
1	TTH, THT, HTT	$\mathbb{P}[\ X = 1\ ] = 3 \times 1/8 = 3/8$	$= 0.375$
2	HHT, HTH, THH	$\mathbb{P}[\ X = 2\ ] = 3 \times 1/8 = 3/8$	$= 0.375$
3	HHH	$\mathbb{P}[\ X = 3\ ] = 1 \times 1/8 = 1/8$	$= 0.125$

Quick Coin Example

A **graphic** is a second way of showing the probability mass function. This is a graphic of the pmf for this problem.



Quick Coin Example

A **formula** is a third way of showing the probability mass function. Such functional representations are handy when the sample space is much larger:

$$\mathbb{P}[\ X = x\ ] = \begin{cases} 0.125 & x = 0 \text{ or } 3 \\ 0.375 & x = 1 \text{ or } 2 \\ 0 & \text{Otherwise} \end{cases}$$

Note that the formula is not unique in its representation. The following also works

$$\mathbb{P}[\ X = x\ ] = \binom{3}{x}(0.5)^x(1-0.5)^{3-x}$$

Population Parameters

Remember that a population parameter is a function of the population. We will usually want to estimate these parameters using our sample statistics. Examples of population parameters include

- mean
- variance
- median

- skew
- success probability
- rate

This section will look at calculating the first three.

Expected Value

The expected value of a distribution (or of a random variable) is the "long-run" average of the distribution's outcomes.

Definition

The **expected value** of a discrete random variable X is equal to the mean of the probability distribution of X and is given by

$$\mathbb{E}\left[X\right] = \sum_{x \in S} x \, \mathbb{P}[\, X = x \,]$$

Variance

The variance of a distribution (or of a random variable) is a measure of uncertainty in each outcome. It has the opposite meaning of *precision*.

Definition

The **variance** of a discrete random variable X is given by

$$\mathbb{V}\left[X\right] = \sum_{x \in S} (x - \mu)^2 \, \mathbb{P}[\, X = x \,]$$

Median

The median of a distribution is an x-value such that at least half is no more than it, and at least half is no *less* than it:

Definition

The **median** of a discrete random variable X is given by

$$\widetilde{X} = \left\{ x \;\middle|\; \mathbb{P}[\, X \leq x \,] \geq 0.50 \text{ and } \mathbb{P}[\, X \geq x \,] \geq 0.50 \right\}$$

Note that the actual definition:

- explains why I hand-waved during the times I discussed the median of a sample
- is much easier in the case of continuous distributions, as it reduces to $\mathbb{P}[\, X \leq \widetilde{x} \,] = 0.50$
- implies that the median is *not* necessarily unique for discrete random variables

Example 1: Three Coins

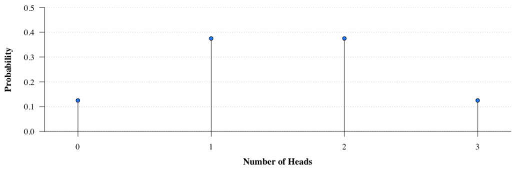
Let us return to our original example, flipping a coin three times. The probability mass function is

$$\mathbb{P}[\, X = x \,] = \begin{cases} 0.125 & x = 0 \text{ or } 3 \\ 0.375 & x = 1 \text{ or } 2 \\ 0 & \text{Otherwise} \end{cases}$$

With this probability mass function, let us calculate the mean, variance, and median.

Example 1: Three Coins

First, here is the probability mass function as a graphic:



From the graphic, what do we expect the mean and median to be?

Example 1: Three Coins

Solution
The mean is defined as

$$\mathbb{E}\left[\, X \,\right] = \sum_{x \in S} x \; \mathbb{P}[\ X = x\ ]$$

Since $S = \{0, 1, 2, 3\}$, the expected number of heads is

$$\begin{aligned} \mathbb{E}\left[\, X \,\right] &= \sum_{x \in S} x \; \mathbb{P}[\ X = x\ ] \\ &= 0(0.125) + 1(0.375) + 2(0.375) + 3(0.125) \\ &= 1.500 \end{aligned}$$

Thus, the expected number of heads on three flips of a fair coin is 1.5. This does not surprise us, right?

Example 1: Three Coins

Solution
The variance is defined as

$$\mathbb{V}\big[\,X\,\big] = \sum_{x \in S} (x-\mu)^2 \; \mathbb{P}[\ X = x\ ]$$

Thus, the variance on the number of heads is

$$\begin{aligned}
\mathbb{V}\big[\,X\,\big] &= \sum_{x \in S} (x-\mu)^2 \; \mathbb{P}[\ X = x\ ] \\
&= (0-1.5)^2(0.125) + (1-1.5)^2(0.375) + \\
&\qquad (2-1.5)^2(0.375) + (3-1.5)^2(0.125) \\
&= 2.25(0.125) + 0.25(0.375) + 0.25(0.375) + 1.25(0.125) \\
&= 0.75 \\
\Longrightarrow SD(X) &= \sqrt{0.75} \approx 0.866
\end{aligned}$$

Example 1: Three Coins

Solution
The median is defined as

$$\widetilde{X} = \big\{ x \mid \mathbb{P}[\ X \leq x\ ] \geq 0.50 \text{ and } \mathbb{P}[\ X \geq x\ ] \geq 0.50 \big\}$$

How do we use this formula?!?! My method is to start low and keep adding until you first get to/over 0.500:

$$\begin{aligned}
X = 0:&\ \ 0.125 \not\geq 0.500 && \text{No success, try the next value of } X \\
X = 1:&\ \ 0.125 + 0.375 \geq 0.500 && \text{Success!!!}
\end{aligned}$$

We have the cumulative probabilities *at least* 0.500, and we are done with the calculations. **Because** the cumulative probabilities *equal* 0.500, both 1 *and* 2 are medians. Technically, the medians are all numbers in the set $1 \leq \widetilde{X} \leq 2$. For the sake of convenience, we will state $\widetilde{X} = 1.5$.

Example 1: Three Coins

R Code
If we understand discrete distributions and how R works, we could use R to get these answers... or to help us get the answers.

- **Mean**

```
x = 0:3
p = c(0.125,0.375,0.375,0.125)
sum(x*p)
```

```
1.5
```


Example 1: Three Coins

R Code
If we understand discrete distributions and how R works, we could use R to get these answers... or to help us get the answers.

- **Variance**

```
x = 0:3
p = c(0.125,0.375,0.375,0.125)
sum( (x-1.5)^2*p )
sqrt(sum( (x-1.5)^2*p ))
```

```
0.75
0.8660254
```

Example 1: Three Coins

R Code

If we understand discrete distributions and how R works, we could use R to get these answers… or to help us get the answers.

- **Median**

```
x = 0:3
p = c(0.125,0.375,0.375,0.125)
cumsum(p)
```

```
0.125 0.500 0.875 1.000
```


Example 2: Ice Hockey

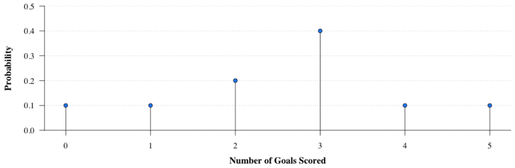
In my STAT 225 course, the course project had my students predict the outcome of an ice hockey game between the Portland Winterhawks and the Prince George Cougars. Together (averaged), they determined that the following was the probability mass function for the number of points scored by the Winterhawks:

Score	0	1	2	3	4	5
Probability	0.1	0.1	0.2	0.4	0.1	0.1

With this information, let us calculate the expected number of goals, the variance, and the median.

Example

First, here is the probability mass function as a graphic:



From the graphic, what do we expect the mean and median to be?

Example 2: Ice Hockey

Solution (Expected Value)
The mean is defined as

$$\mathbb{E}\left[\, X \,\right] = \sum_{x \in S} x \ \mathbb{P}\left[\ X = x\ \right]$$

Given that $S = \{0, 1, 2, 3, 4, 5\}$, the expected number of goals is

$$\begin{aligned}\mathbb{E}\left[\, X \,\right] &= \sum_{x \in S} x \ \mathbb{P}\left[\ X = x\ \right] \\ &= 0(0.1) + 1(0.1) + 2(0.2) + 3(0.4) + 4(0.1) + 5(0.1) \\ &= 2.6\end{aligned}$$

Thus, the expected number of goals to be made by the Winterhawks is 2.6.

Example 2: Ice Hockey

Solution (Variance)
The variance is defined as

$$\mathbb{V}[\,X\,] = \sum_{x \in S} (x-\mu)^2 \; \mathbb{P}[\; X = x \;]$$

Thus, the variance of the number of goals is

$$\begin{aligned}\mathbb{V}[\,X\,] &= \sum_{x \in S} (x-\mu)^2 \; \mathbb{P}[\; X = x \;] \\ &= (0-2.6)^2(0.1) + (1-2.6)^2(0.1) + (2-2.6)^2(0.2) + \\ &\qquad (3-2.6)^2(0.4) + (4-2.6)^2(0.1) + (5-2.6)^2(0.1) \\ &= 6.76(0.1) + 2.56(0.1) + 0.36(0.2) + \\ &\qquad 0.16(0.4) + 1.96(0.1) + 5.76(0.1) \\ &= 1.84\end{aligned}$$

Example 2: Ice Hockey

Aside: The Empirical Rule

Recall the empirical rule from Slidedeck b4. Since the standard deviation is $\sqrt{1.84} \approx 1.356$, we can estimate the probability of the Winterhawks scoring between $\mu - \sigma = 1.244$ and $\mu + \sigma = 3.956$ is about 68%.

Again, remember that the Empirical Rule is *only an approximation*. Since we have the entire probability mass function, we know that the probability of the Winterhawks scoring between 1.244 and 3.956 goals is 0.2 + 0.4 = 60% (not 68%).

Still, this is a rather close estimate, right?

Example 2: Ice Hockey

Solution (Median)
Again, start low and keep adding until you first get to/over 0.500:

$$x = 0: \ 0.1 \not\geq 0.500$$
$$x = 1: \ 0.1 + 0.1 = 0.2 \not\geq 0.500$$
$$x = 2: \ 0.1 + 0.1 + 0.2 = 0.4 \not\geq 0.500$$
$$x = 3: \ 0.1 + 0.1 + 0.2 + 0.4 = 0.8 \geq 0.500 \qquad \text{Success!!!}$$

Thus, a median is 3.

Note: Since the cumulative sum does not *equal* 0.500, the *only* median is 3.

Example 2: Ice Hockey

R Code
If we understand discrete distributions and how R works, we could use R to get these answers… or to help us get the answers.

- **Mean**

```
x = 0:5
p = c(0.1,0.1,0.2,0.4,0.1,0.1)
sum(x*p)
```

```
2.6
```

Example 2: Ice Hockey

R Code
If we understand discrete distributions and how R works, we could use R to get these answers. . . or to help us get the answers.

- **Variance**

```
x = 0:5
p = c(0.1,0.1,0.2,0.4,0.1,0.1)
sum( (x-2.6)^2*p )
sqrt(sum( (x-2.6)^2*p ))
```

```
1.84
1.356466
```


Example 2: Ice Hockey

R Code
If we understand discrete distributions and how R works, we could use R to get these answers. . . or to help us get the answers.

- **Median**

```
x = 0:5
p = c(0.1,0.1,0.2,0.4,0.1,0.1)
cumsum(p)
```

```
0.1 0.2 0.4 0.8 0.9 1.0
```

Today's Objectives

Now that we have concluded this lecture, you should be able to

1. explain what a random variable is
2. understand the difference between discrete and continuous (random) variables
3. know the purpose of the probability mass function (pmf)
4. explain the three requirements for a function to be a pmf
5. calculate probabilities using the pmf
6. determine the sample space of a distribution
7. calculate the expected value and variance of a distribution

Today's R Functions

In this slide deck, we covered the following `R` functions:

- `sum`
- `cumsum`
- `sqrt`

Also, here are six arithmetic operators that may be useful

- `+` addition
- `-` subtraction
- `*` multiplication
- `/` division
- `^` exponentiation
- `:` integer sequence

Supplemental Activities

The following may be of interest to you in terms of today's topics:

- SCA 5a is for some discrete distributions

Note that you can access all Statistical Computing Activities here:
`https://www.kvasaheim.com/courses/stat200/sca/`

In addition to the SCA, **Laboratory Activity B** is helpful for learning how to handle discrete distributions. The lab actually shows the connection between sampling and discrete distributions. It uses three named distributions.
`https://www.kvasaheim.com/courses/stat200/labs/`

Supplemental Readings

The following are some readings that may be of interest to you in terms of understanding discrete distributions:

- Hawkes Learning: Section 5.1
- Intro to Modern Statistics: None
- R for Starters: Appendix A.1