PSY 503: Foundations of Psychological Methods

Lecture 8: Random Variables I

Robin Gomila

Princeton

September 23, 2020

What is a random variable?

- Think about the following random phenomenon: "randomly selecting 2 students in this virtual room"

What is a random variable?

- Think about the following random phenomenon: "randomly selecting 2 students in this virtual room"
- Sample space?

What is a random variable?

- Think about the following random phenomenon: "randomly selecting 2 students in this virtual room"
- Sample space?
- One possible outcome:

What is a random variable?

- Think about the following random phenomenon: "randomly selecting 2 students in this virtual room"
- Sample space?
- One possible outcome: $\omega = \{\text{Tyler}, \text{Stats}\}$

What is a random variable?

- Think about the following random phenomenon: "randomly selecting 2 students in this virtual room"
- Sample space?
- One possible outcome: $\omega = \{\text{Tyler, Stats}\}$
- Another possible outcome: $\omega = \{\text{Sana, Sev}\}$

What is a random variable?

- Think about the following random phenomenon: "randomly selecting 2 students in this virtual room"
- Sample space?
- One possible outcome: $\omega = \{\text{Tyler, Stats}\}$
- Another possible outcome: $\omega = \{\text{Sana, Sev}\}$
- Can this be considered a random variable?

What is a random variable?

- Think about the following random phenomenon: "randomly selecting 2 students in this virtual room"
- Sample space?
- One possible outcome: $\omega = \{\text{Tyler, Stats}\}$
- Another possible outcome: $\omega = \{\text{Sana, Sev}\}$
- Can this be considered a random variable?
    - No.

What is a random variable?

- Think about the following random phenomenon: "randomly selecting 2 students in this virtual room"
- Sample space?
- One possible outcome: $\omega = \{\text{Tyler, Stats}\}$
- Another possible outcome: $\omega = \{\text{Sana, Sev}\}$
- Can this be considered a random variable?
  - No. Random variables are always numeric
  - We operate on random variables using math

Illustration

- For example, a possible random variable (rv) is the number of students in my sample with first letter equal to "S"
- This rv would translate $\{\text{Tyler, Stats}\}$ into the number 1
- This rv would translate $\{\text{Sana, Sev}\}$ into the number 2

Example: Duos and Trios

- Half of the students in a classroom works on their project in pairs (P) and the other half in trios (T)
- We randomly select 2 students in this classroom, and let the random variable W be the number of students who work in pairs

Example: Duos and Trios

- Half of the students in a classroom works on their project in pairs (P) and the other half in trios (T)
- We randomly select 2 students in this classroom, and let the random variable W be the number of students who work in pairs
- Sample space:

Example: Duos and Trios

- Half of the students in a classroom works on their project in pairs (P) and the other half in trios (T)
- We randomly select 2 students in this classroom, and let the random variable W be the number of students who work in pairs
- Sample space: $\Omega = \{\mathrm{TT}, \mathrm{PT}, \mathrm{TP}, \mathrm{PP}\}$

Example: Duos and Trios

- Half of the students in a classroom works on their project in pairs (P) and the other half in trios (T)
- We randomly select 2 students in this classroom, and let the random variable W be the number of students who work in pairs
- Sample space: $\Omega = \{\mathrm{TT}, \mathrm{PT}, \mathrm{TP}, \mathrm{PP}\}$
- Random variable and probability of each outcome

Ω	P(Ω)	W
TT	$\frac{1}{4}$	0
PT	$\frac{1}{4}$	1
TP	$\frac{1}{4}$	1
PP	$\frac{1}{4}$	2

From outcomes to numbers

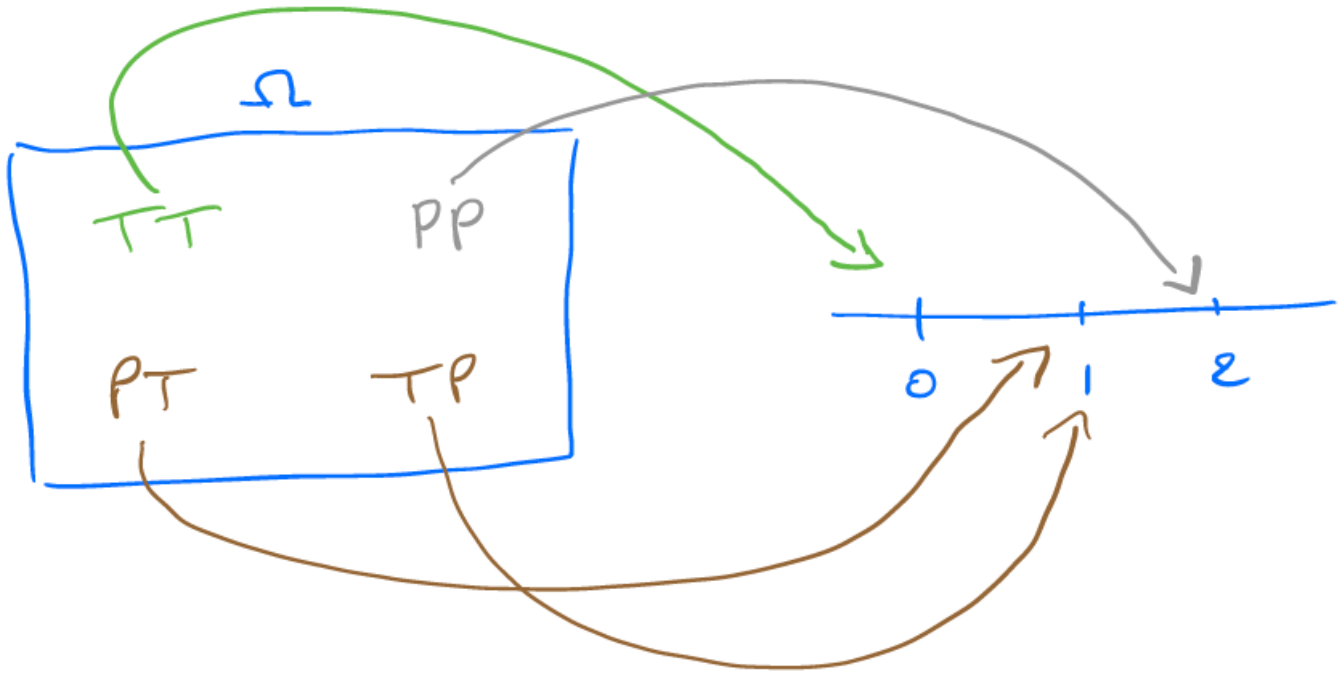
- Random variables are translations of outcomes of a random process into **numbers**

From outcomes to numbers

- Random variables are translations of outcomes of a random process into **numbers**
- Formally, a random variable is defined as a function that maps the sample space Ω of a random generative process into the real line (or into real numbers)

From outcomes to numbers

- Random variables are translations of outcomes of a random process into **numbers**
- Formally, a random variable is defined as a function that maps the sample space Ω of a random generative process into the real line (or into real numbers)



From probabilistic events to random variables

- Let event A be selecting at least one student who is part of a trio.
  - Addition rule:

$$P(A) = P(\text{TT}) + P(\text{PT}) + P(\text{TP}) = \frac{3}{4}$$

From probabilistic events to random variables

- Let event A be selecting at least one student who is part of a trio.
  - Addition rule:

$$P(A) = P(\mathrm{TT}) + P(\mathrm{PT}) + P(\mathrm{TP}) = \frac{3}{4}$$

- Let the random variable Y indicate if at least one of the two selected students is part of a trio
  - Y takes the value 1 if this is the case, and 0 otherwise
  - We write:

$$P(\mathrm{Y} = 1) = \frac{3}{4}$$

Support of a random variable: Definition

The **support** of a random variable is the set of all possible values that a random variable can take.

Distribution functions

- The **distribution** of a random variable X describes the likelihood of the values that X can take
- We will see different distribution functions of random variables
- Earlier, we derived the distribution of a simple rv by directly investigating the underlying sample space

Distribution functions

- We let W be the number of students who work in pairs and found that the distribution of W is

W	$P(W)$
0	$\frac{1}{4}$
1	$\frac{1}{2}$
2	$\frac{1}{4}$

Probability model (sample space) vs. distribution function (rv)

- We are rarely interested in the probability of each outcome from the sample space Ω (e.g., slide 5)
- We are interested in the numeric information that are contained in random variables

Probability model (sample space) vs. distribution function (rv)

- We are rarely interested in the probability of each outcome from the sample space Ω (e.g., slide 5)
- We are interested in the numeric information that are contained in random variables
- Random generative process, sample space, and probabilistic outcomes are always in the background
    - But they provide too much information

Probability model (sample space) vs. distribution function (rv)

- We are rarely interested in the probability of each outcome from the sample space Ω (e.g., slide 5)
- We are interested in the numeric information that are contained in random variables
- Random generative process, sample space, and probabilistic outcomes are always in the background
    - But they provide too much information
- Distribution functions of random variables summarize the **relevant** information for that random generative process

Categories of random variables

- Two types of random variables
  - Discrete
  - Continuous

Categories of random variables

- Two types of random variables
  - Discrete
  - Continuous
- We focus on discrete for now

Discrete random variables: Definition

- Discrete random variables are defined on a range that is a countable set
- i.e., they can only take on a **finite** or **countably infinite** number of different values

Probability Mass Function

- Let X be a discrete rv
- The *probability mass function (PMF)* of X summarizes the probability of each outcome x

Probability Mass Function

- Let X be a discrete rv
- The *probability mass function (PMF)* of X summarizes the probability of each outcome x
- PMF: function p given by

$$p(x) = P(X = x)$$

for all possible values of x

Example: Dessert tonight

Imagine that you started a strict diet a few days ago. You are at a dinner party and realize that your friend made your favorite dessert. You are very tempted and decide to use coin flips to help you make a decision about whether to eat some of that dessert. You will flip a coin three times, the number of times that the flip returns TAILS determines the number of bites that you will have.

Example: Dessert tonight

Before you start flipping the coin, you want to learn more about your chances to have different quantities of dessert tonight. That is, you decide to look at the probability of each possible outcome.

To begin with, you define the random variable X as the number of times a series of three coin flips returns tails (T).

Example: Dessert tonight

- The **support** of X is $\{0, 1, 2, 3\}$, we can write

$$p(x) = \begin{cases} 0 & \text{if (HHH)} \\ 1 & \text{if (HHT) or (HTH) or (THH)} \\ 2 & \text{if (TTH) or (HTT) or (THT)} \\ 3 & \text{if (TTT)} \end{cases} \tag{1}$$

Example: Dessert tonight

- Using the naive definition of probability, we can easily calculate that:

$$p(0) = P(X = 0) = P(\text{CCC}) = \frac{1}{8}$$

$$p(1) = P(X = 1) = P(\text{HHT}) + P(\text{HTH}) + P(\text{THH}) = \frac{3}{8}$$

$$p(2) = P(X = 2) = P(\text{TTH}) + P(\text{HTT}) + P(\text{THT}) = \frac{3}{8}$$

$$p(3) = P(X = 3) = P(\text{TTT}) = \frac{1}{8}$$

Example: Dessert tonight

- Using the naive definition of probability, we can easily calculate that:

$$p(0) = P(X = 0) = P(\text{CCC}) = \frac{1}{8}$$

$$p(1) = P(X = 1) = P(\text{HHT}) + P(\text{HTH}) + P(\text{THH}) = \frac{3}{8}$$

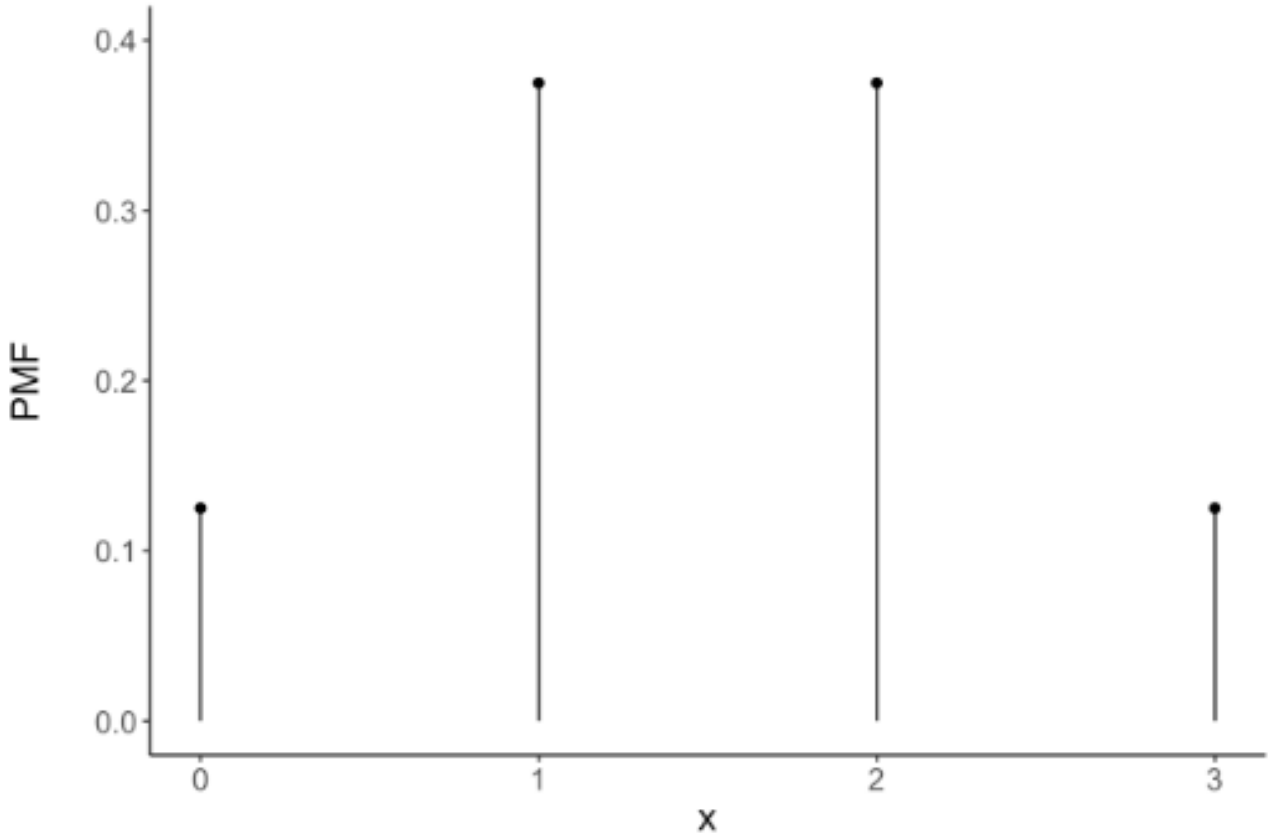
$$p(2) = P(X = 2) = P(\text{TTH}) + P(\text{HTT}) + P(\text{THT}) = \frac{3}{8}$$

$$p(3) = P(X = 3) = P(\text{TTT}) = \frac{1}{8}$$

- This is the content of the PMF of X

PMF of X

- The PMF of X is



PMF of X

- We most often express the PMF in the following way

$$p(x) = \begin{cases} \frac{1}{8} & \text{if } x = 0 \\ \frac{3}{8} & \text{if } x = 1 \\ \frac{3}{8} & \text{if } x = 2 \\ \frac{1}{8} & \text{if } x = 3 \\ 0 & otherwise \end{cases} \tag{2}$$

Property of PMFs

For any value x, we have $0 \leq p(x) \leq 1$, and

$$\sum_{j=1}^{n} p(x_j) = 1 \tag{3}$$

Parametric Discrete Distributions

What are the main characteristics of parametric distributions?

What are the main characteristics of parametric distributions?

- Famous
- Common
- Have their own name

What are the main characteristics of parametric distributions?

- Famous
- Common
- Have their own name
- **Parametric:** they have a pre-calculated PMF that depends entirely on at least one *parameter*
- Once we know the relevant parameter(s), we have all the information we need to calculate the probability of any events, such as the probability that $X = 2$

The Bernoulli distribution

The Bernoulli distribution

- Simplest possible parametric distribution
  - Only one parameter! Generally called θ
- Whenever you see a variable that is binary
  - i.e., that can take on only two values: 0 and 1

The Bernoulli distribution

- θ indicates the "probability of success"
  - Probability that the random variable returns 1

The Bernoulli distribution

- θ indicates the "probability of success"
  - Probability that the random variable returns 1
- Formally, for any Bernoulli rv, we have:

$$P(X = 1) = \theta$$

$$P(X = 0) = 1 - \theta$$

The Bernoulli distribution

- θ indicates the "probability of success"
  - Probability that the random variable returns 1
- Formally, for any Bernoulli rv, we have:

$$P(X = 1) = \theta$$

$$P(X = 0) = 1 - \theta$$

- If you know θ, you know everything:
  - e.g., mean, median, variance, standard deviation, mode (more soon!)

The Bernoulli distribution

- θ indicates the "probability of success"
  - Probability that the random variable returns 1
- Formally, for any Bernoulli rv, we have:

$$P(X = 1) = \theta$$

$$P(X = 0) = 1 - \theta$$

- If you know θ, you know everything:
  - e.g., mean, median, variance, standard deviation, mode (more soon!)
- We write $X \sim Bern(\theta)$

Example: Trendy Dining in NYC

Imagine that you manage a trendy dining in New York that has **capacity for 30 tables.** You are fully booked until the end of the year, but you are considering overbooking because every night, some reservations do not show up. Specifically, **you know that on average, only 90% of the reservations show up on a given night.**

Example: Trendy Dining in NYC

Imagine that you manage a trendy dining in New York that has **capacity for 30 tables.** You are fully booked until the end of the year, but you are considering overbooking because every night, some reservations do not show up. Specifically, **you know that on average, only 90% of the reservations show up on a given night.**

Let X_i be a random variable that indicates whether reservation i showed up on a given night.

Example: Trendy Dining in NYC

Imagine that you manage a trendy dining in New York that has **capacity for 30 tables.** You are fully booked until the end of the year, but you are considering overbooking because every night, some reservations do not show up. Specifically, **you know that on average, only 90% of the reservations show up on a given night.**

Let X_i be a random variable that indicates whether reservation i showed up on a given night.

X_i is a Bernoulli random variable because it can only take on the value 0 (if a reservation did not show up) and 1 (if a reservation did show up).

Example: Trendy Dining in NYC

- The parameter θ is the probability of "success":
  - i.e., probability that a reservation shows up.
- We have:
  - $\theta =$

Example: Trendy Dining in NYC

- The parameter θ is the probability of "success":
  - i.e., probability that a reservation shows up.
- We have:
  - $\theta = 0.90$

Example: Trendy Dining in NYC

- The parameter θ is the probability of "success":
  - i.e., probability that a reservation shows up.
- We have:
  - $\theta = 0.90$
  - $P(X = 1) =$

Example: Trendy Dining in NYC

- The parameter θ is the probability of "success":
  - i.e., probability that a reservation shows up.
- We have:
  - $\theta = 0.90$
  - $P(X = 1) = \theta = 0.90$

Example: Trendy Dining in NYC

- The parameter θ is the probability of "success":
  - i.e., probability that a reservation shows up.
- We have:
  - $\theta = 0.90$
  - $P(X = 1) = \theta = 0.90$
  - $P(X = 0) =$

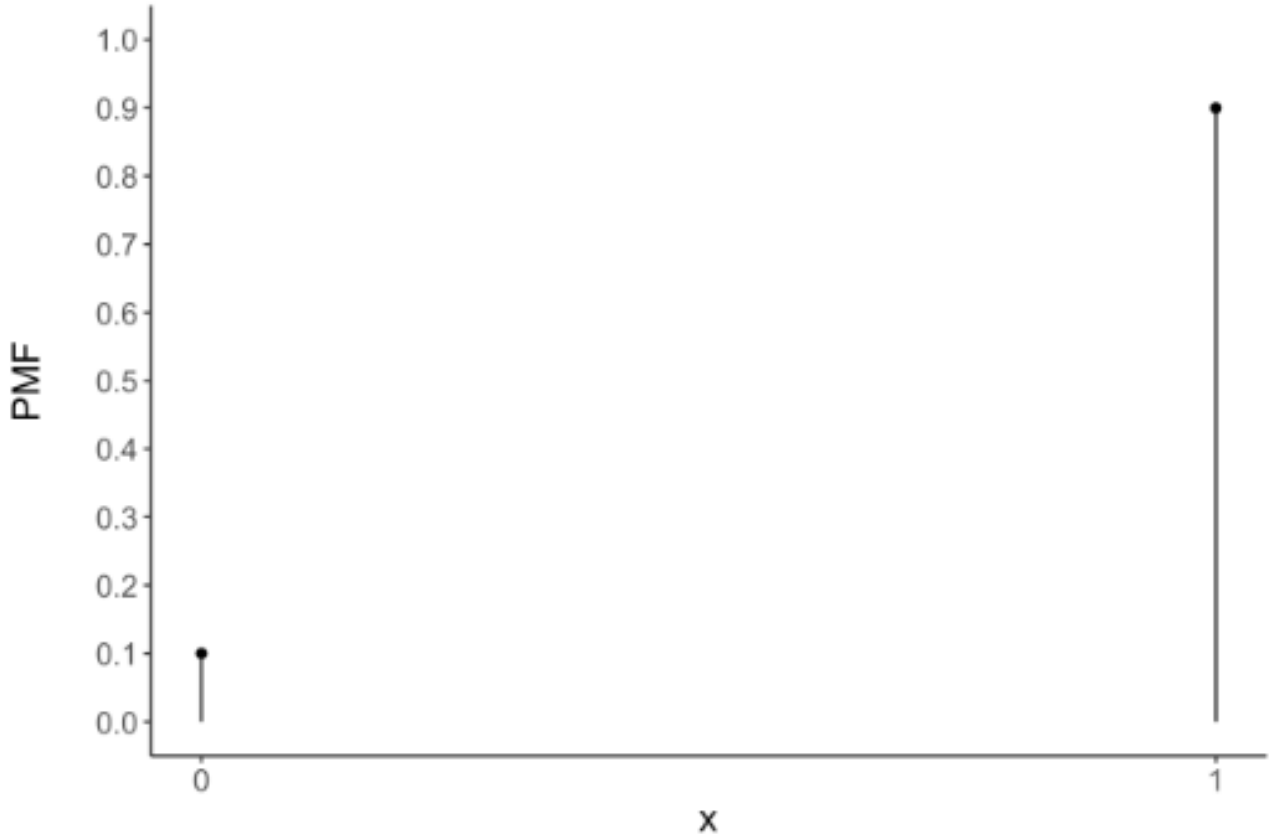
Example: Trendy Dining in NYC

- The parameter θ is the probability of "success":
  - i.e., probability that a reservation shows up.
- We have:
  - $\theta = 0.90$
  - $P(X = 1) = \theta = 0.90$
  - $P(X = 0) = 1 - \theta = 0.10$

Trendy Dining: PMF

$$p(x) = \begin{cases} 0.10 & \text{if } x = 0 \\ 0.90 & \text{if } x = 1 \\ 0 & otherwise \end{cases} \tag{4}$$

Trendy Dining: PMF



Trendy Dining: New strategies

Imagine that you decide to adopt a new strategy: accepting more reservation than you can host. To test the water, you will accept a total of 34 reservations for August 21 and 35 reservations for August 22. Let's use R to simulate what happened on these two nights.

Trendy Dining: Simulations

```
set.seed(0821)

august21_night <- sample(0:1,
                         34,
                         replace = TRUE,
                         prob = c(.10, .90))

head(august21_night, 10)
```

```
##  [1] 1 1 1 0 1 1 1 0 1 1
```

```
sum(august21_night)
```

```
## [1] 29
```

Trendy Dining: Simulations

YAY! 29 tables showed up on August 21.

Let's see what happened on the next day.

Trendy Dining: Simulations

```
set.seed(0822)

august22_night <- sample(0:1,
                         35,
                         replace = TRUE,
                         prob = c(.10, .90))

head(august22_night, 10)
```

```
##  [1] 1 0 1 1 1 1 1 1 1 0
```

```
sum(august22_night)
```

```
## [1] 31
```

Trendy Dining: Simulations

Terrible! You weren't able to accommodate all of your reservations on that night. Not good for your your yelp reviews!

Trendy Dining: What to do on August 23?

- How lucky did you get on August 21? How unlucky did you get on August 22?
- How often will it be the case that if you book a certain number (e.g., 4) of extra tables every night, you will not be able to serve all of those that show up?
- We need to learn about a related but different parametric distribution: the **binomial distribution**

Binomial Distribution

- When we are interested in the number of *successes* in a series of Bernoulli trials
- Two *parameters*: θ and n
- θ: Probability of success
- n: number of Bernoulli trials included in X_i

Trendy Dining: Binomial parameters

Suppose you are interested in the probability that exactly 30 people will show up at the restaurant (i.e., 30 successes) when you accepted 34 reservations. You know that the proportion of reservation that actually show up on a given night is .90.

What is θ?

What is n?

Trendy Dining: Binomial parameters

Suppose you are interested in the probability that exactly 30 people will show up at the restaurant (i.e., 30 successes) when you accepted 34 reservations. You know that the proportion of reservation that actually show up on a given night is .90.

$\theta = .90$

Trendy Dining: Binomial parameters

Suppose you are interested in the probability that exactly 30 people will show up at the restaurant (i.e., 30 successes) when you accepted 34 reservations. You know that the proportion of reservation that actually show up on a given night is .90.

$\theta = .90$

$n = 34$

Trendy Dining: Binomial parameters

Suppose you are interested in the probability that exactly 30 people will show up at the restaurant (i.e., 30 successes) when you accepted 34 reservations. You know that the proportion of reservation that actually show up on a given night is .90.

$\theta = .90$

$n = 34$

We write $X \sim Bin(n, \theta)$

Trendy Dining: Binomial parameters

Suppose you are interested in the probability that exactly 30 people will show up at the restaurant (i.e., 30 successes) when you accepted 34 reservations. You know that the proportion of reservation that actually show up on a given night is .90.

$\theta = .90$

$n = 34$

We write $X \sim Bin(n, \theta)$

In this example, we have $X \sim Bin(34, .90)$

PMF of $X \sim Bin(n, \theta)$

The PMF of X is

$$\frac{n!}{(n-x)!x!}\theta^x(1-\theta)^{n-x} \tag{5}$$

Trendy Dining: PMF

Let X be the number of "successes" from a series of $n = 34$ Bernoulli trials with probability of success $\theta = 0.90$. We have all of the parameters for this Binomial random variable X, and we can use Equation 5 to derive $P(X = 30)$:

Trendy Dining: PMF

Let X be the number of "successes" from a series of $n = 34$ Bernoulli trials with probability of success $\theta = 0.90$. We have all of the parameters for this Binomial random variable X, and we can use Equation 5 to derive $P(X = 30)$:

$$P(X = 30) = \frac{34!}{(34 - 30)!30!} 0.90^{30} (1 - 0.90)^{34-30} = 0.19659$$

Trendy Dining: PMF using R

- Use the dbinom() function

```
dbinom(x = 30,
       size = 34,
       prob = 0.9)
```

```
## [1] 0.1965932
```

Trendy Dining: PMF using R

We can also use the function dbinom() to calculate the probability of multiple values x. For instance, if we wanted the probability of each possible outcomes, we could write:

```
x_vector <- 0:34
all_probs <- dbinom(x = x_vector,
                    size = 34,
                    prob = 0.9)
```

Trendy Dining: PMF using R