College of Media, Communication and Information
UNIVERSITY OF COLORADO BOULDER

Probability Distributions: Discrete

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

FEBRUARY 19, 2017

Refresher: Random variables

- Random variables take on values in a *sample space*.
- This week we will focus on *discrete* random variables:
  - Coin flip: $\{H, T\}$
  - Number of times a coin lands heads after N flips: $\{0, 1, 2, \ldots, N\}$
  - Number of words in a document: Positive integers $\{1, 2, \ldots\}$
- Reminder: we denote the random variable with a capital letter; denote a outcome with a lower case letter.
  - E.g., X is a coin flip, x is the value (H or T) of that coin flip.

Refresher: Discrete distributions

- A discrete distribution assigns a probability to every possible outcome in the sample space
- For example, if X is a coin flip, then

$$\begin{aligned} P(X=H) &= 0.5 \\ P(X=T) &= 0.5 \end{aligned}$$

- Probabilities have to be greater than or equal to 0 and probabilities over the entire sample space must sum to one

$$\sum_x P(X=x) = 1$$

Mathematical Conventions

0!

If $n! = n \cdot (n-1)!$ then $0! = 1$ if definition holds for $n > 0$.

n^0

Example for 3:

$$3^2 = 9 \tag{1}$$

$$3^1 = 3 \tag{2}$$

$$3^{-1} = \frac{1}{3} \tag{3}$$

Mathematical Conventions

0!

If $n! = n \cdot (n-1)!$ then $0! = 1$ if definition holds for $n > 0$.

n^0

Example for 3:

$$3^2 = 9 \quad (1)$$
$$3^1 = 3 \quad (2)$$
$$3^0 = 1 \quad (3)$$
$$3^{-1} = \frac{1}{3} \quad (4)$$

Today: Types of discrete distributions

- There are many different types of discrete distributions, with different definitions.
- Today we'll look at the most common discrete distributions.
  - And we'll introduce the concept of *parameters*.
- These discrete distributions (along with the continuous distributions next) are fundamental

Probability Distributions: Discrete

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

FEBRUARY 19, 2017

Bernoulli distribution

- A distribution over a sample space with two values: $\{0, 1\}$
  - Interpretation: 1 is “success”; 0 is “failure”
  - Example: coin flip (we let 1 be “heads” and 0 be “tails”)
- A Bernoulli distribution can be defined with a table of the two probabilities:
  - X denotes the outcome of a coin flip:

$$\begin{aligned} P(X=0) &= 0.5 \\ P(X=1) &= 0.5 \end{aligned}$$

  - X denotes whether or not a TV is defective:

$$\begin{aligned} P(X=0) &= 0.995 \\ P(X=1) &= 0.005 \end{aligned}$$

Bernoulli distribution

- Do we need to write out both probabilities?

$$\begin{aligned} P(X=0) &= 0.995 \\ P(X=1) &= 0.005 \end{aligned}$$

- What if I only told you $P(X=1)$? Or $P(X=0)$?

Bernoulli distribution

- Do we need to write out both probabilities?

$$
\begin{aligned}
P(X=0) &= 0.995 \\
P(X=1) &= 0.005
\end{aligned}
$$

- What if I only told you $P(X=1)$? Or $P(X=0)$?

$$
\begin{aligned}
P(X=0) &= 1-P(X=1) \\
P(X=1) &= 1-P(X=0)
\end{aligned}
$$

- We only need one probability to define a Bernoulli distribution
  - Usually the probability of success, $P(X=1)$.

Another way of writing the Bernoulli distribution:

- Let θ denote the probability of success ($0 \leq \theta \leq 1$).

$$\begin{aligned} P(X=0) &= 1-\theta \\ P(X=1) &= \theta \end{aligned}$$

- An even more compact way to write this:

$$P(X=x) = \theta^x (1-\theta)^{1-x}$$

  - This is called a *probability mass function*.

Probability mass functions

- A probability mass function (PMF) is a function that assigns a probability to every outcome of a discrete random variable X.
  - Notation: $f(x) = P(X = x)$
- Compact definition
- Example: PMF for Bernoulli random variable $X \in \{0, 1\}$

$$f(x) = \theta^x (1 - \theta)^{1-x}$$

- In this example, θ is called a *parameter*.

Parameters

- Define the probability mass function
- *Free parameters* not constrained by the PMF.
- For example, the Bernoulli PMF could be written with two parameters:

$$f(x) = \theta_1^x \theta_2^{1-x}$$

  But $\theta_2 \equiv 1 - \theta_1$ ...only 1 free parameter.
- The *complexity* $\approx$ number of free parameters. Simpler models have fewer parameters.

Sampling from a Bernoulli distribution

- How to randomly generate a value distributed according to a Bernoulli distribution?
- Algorithm:
  1. Randomly generate a number between 0 and 1
     r = random(0, 1)
  2. If $r < \theta$, return success
     Else, return failure

College of Media, Communication and Information
UNIVERSITY OF COLORADO BOULDER

Probability Distributions: Discrete

INFO-2301: Quantitative Reasoning 2
Michael Paul and Jordan Boyd-Graber

FEBRUARY 19, 2017

Binomial distribution

- Bernoulli: distribution over two values (success or failure) from a single event
- **binomial**: number of successes from *multiple* Bernoulli events
- Examples:
  - The number of times "heads" comes up after flipping a coin 10 times
  - The number of defective TVs in a line of 10,000 TVs
- Important: each Bernoulli event is assumed to be independent
- Notation: let X be a random variable that describes the number of successes out of N trials.
  - The possible values of X are integers from 0 to N: $\{0, 1, 2, \dots, N\}$

Binomial distribution

- Suppose we flip a coin 3 times. There are 8 possible outcomes:

$$
\begin{aligned}
P(HHH) &= P(H)P(H)P(H) &= 0.125 \\
P(HHT) &= P(H)P(H)P(T) &= 0.125 \\
P(HTH) &= P(H)P(T)P(H) &= 0.125 \\
P(HTT) &= P(H)P(T)P(T) &= 0.125 \\
P(THH) &= P(T)P(H)P(H) &= 0.125 \\
P(THT) &= P(T)P(H)P(T) &= 0.125 \\
P(TTH) &= P(T)P(T)P(H) &= 0.125 \\
P(TTT) &= P(T)P(T)P(T) &= 0.125
\end{aligned}
$$

- What is the probability of landing heads x times during these 3 flips?

Binomial distribution

- What is the probability of landing heads x times during these 3 flips?
- 0 times:
  - $P(TTT) = 0.125$
- 1 time:
  - $P(HTT) + P(THT) + P(TTH) = 0.375$
- 2 times:
  - $P(HHT) + P(HTH) + P(THH) = 0.375$
- 3 times:
  - $P(HHH) = 0.125$

Binomial distribution

- The probability mass function for the binomial distribution is:

$$f(x) = \underbrace{\binom{N}{x}}_{\text{“}N\text{ choose }x\text{”}} \theta^x (1-\theta)^{N-x}$$

- Like the Bernoulli, the binomial parameter θ is the probability of success from one event.
- Binomial has second parameter *N*: number of trials.
- The PMF important: difficult to figure out the entire distribution by hand.

Aside: Binomial coefficients

- The expression $\binom{n}{k}$ is called a *binomial coefficient*.
  - Also called a *combination* in combinatorics.
- $\binom{n}{k}$ is the number of ways to choose k elements from a set of n elements.
- For example, the number of ways to choose 2 heads from 3 coin flips:
  HHT, HTH, THH
  $\binom{3}{2} = 3$
- Formula:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

1
1 1
1 2 1
1 3 3 1
1 4 6 4 1
1 5 10 10 5 1

Pascal's triangle depicts the values of $\binom{n}{k}$.

Bernoulli vs Binomial

- A Bernoulli distribution is a special case of the binomial distribution when $N = 1$.
- For this reason, sometimes the term binomial is used to refer to a Bernoulli random variable.

Example

- Probability that a coin lands heads *at least* once during 3 flips?

Example

- Probability that a coin lands heads *at least* once during 3 flips?

$$P(X \geq 1)$$

Example

- Probability that a coin lands heads *at least* once during 3 flips?

$$
\begin{aligned}
P(X \geq 1) &= P(X=1) + P(X=2) + P(X=3) \\
&= 0.375 + 0.375 + 0.125 = 0.875
\end{aligned}
$$

College of Media, Communication and Information
UNIVERSITY OF COLORADO BOULDER

Probability Distributions: Discrete

INFO-2301: Quantitative Reasoning 2
Michael Paul and Jordan Boyd-Graber

FEBRUARY 19, 2017

Categorical distribution

- Recall: the Bernoulli distribution is a distribution over two values (success or failure)
- **categorical** distribution generalizes Bernoulli distribution over any number of values
  - Rolling a die
  - Selecting a card from a deck
- AKA *discrete* distribution.
  - Most general type of discrete distribution
  - specify all (but one) of the probabilities in the distribution
  - rather than the probabilities being determined by the probability mass function.

Categorical distribution

- If the categorical distribution is over K possible outcomes, then the distribution has K parameters.
- We will denote the parameters with a K-dimensional vector $\vec{\theta}$.
- The probability mass function can be written as:

$$f(x) = \prod_{k=1}^{K} \theta_k^{[x=k]}$$

  where the expression $[x = k]$ evaluates to 1 if the statement is true and 0 otherwise.
    - All this really says is that the probability of outcome x is equal to θ_x.
- The number of *free parameters* is $K-1$, since if you know $K-1$ of the parameters, the Kth parameter is constrained to sum to 1.

Categorical distribution

- Example: the roll of a (unweighted) die

$$
\begin{aligned}
P(X=1) &= \tfrac{1}{6} \\
P(X=2) &= \tfrac{1}{6} \\
P(X=3) &= \tfrac{1}{6} \\
P(X=4) &= \tfrac{1}{6} \\
P(X=5) &= \tfrac{1}{6} \\
P(X=6) &= \tfrac{1}{6}
\end{aligned}
$$

- If all outcomes have equal probability, this is called the *uniform* distribution.
- General notation: $P(X=x)=\theta_x$

Sampling from a categorical distribution

- How to randomly select a value distributed according to a categorical distribution?
- The idea is similar to randomly selected a Bernoulli-distributed value.
- Algorithm:
  1. Randomly generate a number between 0 and 1
     r = random(0, 1)
  2. For $k = 1, \ldots, K$:
     - Return smallest r s.t. $r < \sum_{i=1}^{k} \theta_k$

Sampling from a categorical distribution

- Example: simulating the roll of a die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.166667 \\
P(X=2) &= \theta_2 &= 0.166667 \\
P(X=3) &= \theta_3 &= 0.166667 \\
P(X=4) &= \theta_4 &= 0.166667 \\
P(X=5) &= \theta_5 &= 0.166667 \\
P(X=6) &= \theta_6 &= 0.166667
\end{aligned}
$$

Sampling from a categorical distribution

- Example: simulating the roll of a die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.166667 \\
P(X=2) &= \theta_2 &= 0.166667 \\
P(X=3) &= \theta_3 &= 0.166667 \\
P(X=4) &= \theta_4 &= 0.166667 \\
P(X=5) &= \theta_5 &= 0.166667 \\
P(X=6) &= \theta_6 &= 0.166667
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.452383$

Sampling from a categorical distribution

- Example: simulating the roll of a die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.166667 \\
P(X=2) &= \theta_2 &= 0.166667 \\
P(X=3) &= \theta_3 &= 0.166667 \\
P(X=4) &= \theta_4 &= 0.166667 \\
P(X=5) &= \theta_5 &= 0.166667 \\
P(X=6) &= \theta_6 &= 0.166667
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.452383$

$r < \theta_1$?

Sampling from a categorical distribution

- Example: simulating the roll of a die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.166667 \\
P(X=2) &= \theta_2 &= 0.166667 \\
P(X=3) &= \theta_3 &= 0.166667 \\
P(X=4) &= \theta_4 &= 0.166667 \\
P(X=5) &= \theta_5 &= 0.166667 \\
P(X=6) &= \theta_6 &= 0.166667
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.452383$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?

Sampling from a categorical distribution

- Example: simulating the roll of a die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.166667 \\
P(X=2) &= \theta_2 &= 0.166667 \\
P(X=3) &= \theta_3 &= 0.166667 \\
P(X=4) &= \theta_4 &= 0.166667 \\
P(X=5) &= \theta_5 &= 0.166667 \\
P(X=6) &= \theta_6 &= 0.166667
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.452383$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?
$r < \theta_1 + \theta_2 + \theta_3$?

Sampling from a categorical distribution

- Example: simulating the roll of a die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.166667 \\
P(X=2) &= \theta_2 &= 0.166667 \\
P(X=3) &= \theta_3 &= 0.166667 \\
P(X=4) &= \theta_4 &= 0.166667 \\
P(X=5) &= \theta_5 &= 0.166667 \\
P(X=6) &= \theta_6 &= 0.166667
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.452383$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?
$r < \theta_1 + \theta_2 + \theta_3$?

- Return $X = 3$

Sampling from a categorical distribution

- Example: simulating the roll of a die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.166667 \\
P(X=2) &= \theta_2 &= 0.166667 \\
P(X=3) &= \theta_3 &= 0.166667 \\
P(X=4) &= \theta_4 &= 0.166667 \\
P(X=5) &= \theta_5 &= 0.166667 \\
P(X=6) &= \theta_6 &= 0.166667
\end{aligned}
$$

Sampling from a categorical distribution

- Example: simulating the roll of a die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.166667 \\
P(X=2) &= \theta_2 &= 0.166667 \\
P(X=3) &= \theta_3 &= 0.166667 \\
P(X=4) &= \theta_4 &= 0.166667 \\
P(X=5) &= \theta_5 &= 0.166667 \\
P(X=6) &= \theta_6 &= 0.166667
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.117544$

Sampling from a categorical distribution

- Example: simulating the roll of a die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.166667 \\
P(X=2) &= \theta_2 &= 0.166667 \\
P(X=3) &= \theta_3 &= 0.166667 \\
P(X=4) &= \theta_4 &= 0.166667 \\
P(X=5) &= \theta_5 &= 0.166667 \\
P(X=6) &= \theta_6 &= 0.166667
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.117544$

$r < \theta_1$?

Sampling from a categorical distribution

- Example: simulating the roll of a die

$$\begin{aligned}
P(X=1) &= \theta_1 &= 0.166667 \\
P(X=2) &= \theta_2 &= 0.166667 \\
P(X=3) &= \theta_3 &= 0.166667 \\
P(X=4) &= \theta_4 &= 0.166667 \\
P(X=5) &= \theta_5 &= 0.166667 \\
P(X=6) &= \theta_6 &= 0.166667
\end{aligned}$$

Random number in $(0,1)$:
$r = 0.117544$

$r < \theta_1$?

- Return $X = 1$

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.01 \\
P(X=2) &= \theta_2 &= 0.01 \\
P(X=3) &= \theta_3 &= 0.01 \\
P(X=4) &= \theta_4 &= 0.01 \\
P(X=5) &= \theta_5 &= 0.01 \\
P(X=6) &= \theta_6 &= 0.95
\end{aligned}
$$

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$\begin{aligned} P(X=1) &= \theta_1 &= 0.01 \\ P(X=2) &= \theta_2 &= 0.01 \\ P(X=3) &= \theta_3 &= 0.01 \\ P(X=4) &= \theta_4 &= 0.01 \\ P(X=5) &= \theta_5 &= 0.01 \\ P(X=6) &= \theta_6 &= 0.95 \end{aligned}$$

Random number in $(0,1)$:
$r = 0.209581$

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.01 \\
P(X=2) &= \theta_2 &= 0.01 \\
P(X=3) &= \theta_3 &= 0.01 \\
P(X=4) &= \theta_4 &= 0.01 \\
P(X=5) &= \theta_5 &= 0.01 \\
P(X=6) &= \theta_6 &= 0.95
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.209581$

$r < \theta_1$?

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.01 \\
P(X=2) &= \theta_2 &= 0.01 \\
P(X=3) &= \theta_3 &= 0.01 \\
P(X=4) &= \theta_4 &= 0.01 \\
P(X=5) &= \theta_5 &= 0.01 \\
P(X=6) &= \theta_6 &= 0.95
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.209581$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.01 \\
P(X=2) &= \theta_2 &= 0.01 \\
P(X=3) &= \theta_3 &= 0.01 \\
P(X=4) &= \theta_4 &= 0.01 \\
P(X=5) &= \theta_5 &= 0.01 \\
P(X=6) &= \theta_6 &= 0.95
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.209581$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?
$r < \theta_1 + \theta_2 + \theta_3$?

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.01 \\
P(X=2) &= \theta_2 &= 0.01 \\
P(X=3) &= \theta_3 &= 0.01 \\
P(X=4) &= \theta_4 &= 0.01 \\
P(X=5) &= \theta_5 &= 0.01 \\
P(X=6) &= \theta_6 &= 0.95
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.209581$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?
$r < \theta_1 + \theta_2 + \theta_3$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4$?

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.01 \\
P(X=2) &= \theta_2 &= 0.01 \\
P(X=3) &= \theta_3 &= 0.01 \\
P(X=4) &= \theta_4 &= 0.01 \\
P(X=5) &= \theta_5 &= 0.01 \\
P(X=6) &= \theta_6 &= 0.95
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.209581$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?
$r < \theta_1 + \theta_2 + \theta_3$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5$?

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.01 \\
P(X=2) &= \theta_2 &= 0.01 \\
P(X=3) &= \theta_3 &= 0.01 \\
P(X=4) &= \theta_4 &= 0.01 \\
P(X=5) &= \theta_5 &= 0.01 \\
P(X=6) &= \theta_6 &= 0.95
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.209581$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?
$r < \theta_1 + \theta_2 + \theta_3$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5 + \theta_6$?

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$
\begin{aligned}
P(X=1) &= \theta_1 &= 0.01 \\
P(X=2) &= \theta_2 &= 0.01 \\
P(X=3) &= \theta_3 &= 0.01 \\
P(X=4) &= \theta_4 &= 0.01 \\
P(X=5) &= \theta_5 &= 0.01 \\
P(X=6) &= \theta_6 &= 0.95
\end{aligned}
$$

Random number in $(0,1)$:
$r = 0.209581$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?
$r < \theta_1 + \theta_2 + \theta_3$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5 + \theta_6$?

- Return $X = 6$

Sampling from a categorical distribution

- Example 2: rolling a *biased* die

$$\begin{aligned} P(X=1) &= \theta_1 &= 0.01 \\ P(X=2) &= \theta_2 &= 0.01 \\ P(X=3) &= \theta_3 &= 0.01 \\ P(X=4) &= \theta_4 &= 0.01 \\ P(X=5) &= \theta_5 &= 0.01 \\ P(X=6) &= \theta_6 &= 0.95 \end{aligned}$$

Random number in $(0,1)$:
$r = 0.209581$

$r < \theta_1$?
$r < \theta_1 + \theta_2$?
$r < \theta_1 + \theta_2 + \theta_3$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5$?
$r < \theta_1 + \theta_2 + \theta_3 + \theta_4 + \theta_5 + \theta_6$?

- Return $X = 6$

- We will always return $X = 6$ unless our random number $r < 0.05$.
  - 6 is the most probable outcome

College of Media, Communication and Information
UNIVERSITY OF COLORADO BOULDER

Probability Distributions: Discrete

INFO-2301: Quantitative Reasoning 2

Michael Paul and Jordan Boyd-Graber

FEBRUARY 19, 2017

Multinomial distribution

- Recall: the binomial distribution is the number of successes from multiple Bernoulli success/fail events
- The **multinomial** distribution is the number of different outcomes from multiple *categorical* events
  - It is a generalization of the binomial distribution to more than two possible outcomes
  - As with the binomial distribution, each categorical event is assumed to be independent
  - **Bernoulli : binomial :: categorical : multinomial**
- Examples:
  - The number of times each face of a die turned up after 50 rolls
  - The number of times each suit is drawn from a deck of cards after 10 draws

Multinomial distribution

- Notation: let $\vec{X}$ be a vector of length K, where X_k is a random variable that describes the number of times that the kth value was the outcome out of N categorical trials.
    - The possible values of each X_k are integers from 0 to N
    - All X_k values must sum to N: $\sum_{k=1}^{K} X_k = N$
- Example: if we roll a die 10 times, suppose it comes up with the following values:

$$\vec{X} = < 1, 0, 3, 2, 1, 3 >$$

$$X_1 = 1$$
$$X_2 = 0$$
$$X_3 = 3$$
$$X_4 = 2$$
$$X_5 = 1$$
$$X_6 = 3$$

- The multinomial distribution is a *joint* distribution over multiple random variables: $P(X_1, X_2, \ldots, X_K)$

Multinomial distribution

- Suppose we roll a die 3 times. There are 216 (6^3) possible outcomes:

$$
\begin{aligned}
P(111) &= P(1)P(1)P(1) &= 0.00463 \\
P(112) &= P(1)P(1)P(2) &= 0.00463 \\
P(113) &= P(1)P(1)P(3) &= 0.00463 \\
P(114) &= P(1)P(1)P(4) &= 0.00463 \\
P(115) &= P(1)P(1)P(5) &= 0.00463 \\
P(116) &= P(1)P(1)P(6) &= 0.00463 \\
\ldots & \qquad \ldots & \ldots \\
P(665) &= P(6)P(6)P(5) &= 0.00463 \\
P(666) &= P(6)P(6)P(6) &= 0.00463
\end{aligned}
$$

- What is the probability of a particular vector of counts after 3 rolls?

Multinomial distribution

- What is the probability of a particular vector of counts after 3 rolls?
- Example 1: $\vec{X} =< 0, 1, 0, 0, 2, 0 >$

Multinomial distribution

- What is the probability of a particular vector of counts after 3 rolls?
- Example 1: $\vec{X} = < 0, 1, 0, 0, 2, 0 >$
  - $P(\vec{X}) = P(255) + P(525) + P(552) = 0.01389$

Multinomial distribution

- What is the probability of a particular vector of counts after 3 rolls?
- Example 1: $\vec{X} = < 0, 1, 0, 0, 2, 0 >$
  - $P(\vec{X}) = P(255) + P(525) + P(552) = 0.01389$
- Example 2: $\vec{X} = < 0, 0, 1, 1, 1, 0 >$

Multinomial distribution

- What is the probability of a particular vector of counts after 3 rolls?
- Example 1: $\vec{X} =< 0, 1, 0, 0, 2, 0 >$
  - $P(\vec{X}) = P(255) + P(525) + P(552) = 0.01389$
- Example 2: $\vec{X} =< 0, 0, 1, 1, 1, 0 >$
  - $P(\vec{X}) = P(345) + P(354) + P(435) + P(453) + P(534) + P(543) = 0.02778$

Multinomial distribution

- The probability mass function for the multinomial distribution is:

$$f(\vec{x}) = \underbrace{\frac{N!}{\prod_{k=1}^{K} x_k!}}_{\text{Generalization of binomial coefficient}} \prod_{k=1}^{K} \theta_k^{x_k}$$

- Like categorical distribution, multinomial has a K-length parameter vector $\vec{\theta}$ encoding the probability of each outcome.
- Like binomial, the multinomial distribution has a additional parameter N, which is the number of events.

Multinomial distribution: summary

- Categorical distribution is multinomial when $N = 1$.
- Sampling from a multinomial: same code repeated N times.
  - Remember that each categorical trial is independent.
  - Question: Does this mean the count values (i.e., each X_1, X_2, etc.) are independent?

Multinomial distribution: summary

- Categorical distribution is multinomial when $N = 1$.
- Sampling from a multinomial: same code repeated N times.
    - Remember that each categorical trial is independent.
    - Question: Does this mean the count values (i.e., each X_1, X_2, etc.) are independent?
        - No! If $N = 3$ and $X_1 = 2$, then X_2 can be no larger than 1 (must sum to N).

Multinomial distribution: summary

- Categorical distribution is multinomial when $N = 1$.
- Sampling from a multinomial: same code repeated N times.
  - Remember that each categorical trial is independent.
  - Question: Does this mean the count values (i.e., each X_1, X_2, etc.) are independent?
    - No! If $N = 3$ and $X_1 = 2$, then X_2 can be no larger than 1 (must sum to N).
- Remember this analogy:
  - **Bernoulli : binomial :: categorical : multinomial**