Discrete Random Variables

Probability Theory: Foundation for Data Science

with **Anne Dougherty**

Data Science
UNIVERSITY OF COLORADO **BOULDER**

Random Variables

At the end of this module, students should be able to

- Define a discrete random variable and give examples of a probability mass function and a cumulative distribution function.
- Calculate probabilities of Bernoulli, Binomial, Geometric, Negative Binomial, and Geometric random variables.
- Calculate the expectation and variance.

Review

So far, we've learned about:

- Experiments, events, sample spaces
- Axioms of Probability and their consequences
- Conditional Probability and Bayes Theorem
- Permutations and Combinations
- Independent Events

We need a framework for more easily representing events — Random Variables

Random Variables

Definition: A random variable (rv) is a function that maps events (from the sample space S) to the real numbers.

Random variables can be **discrete** or **continuous**, or sometimes a mixture of the two.

A rv is **discrete** if its set of possible values is discrete.

Example: Flip a fair coin 100 times. Let Y be the number of heads in the 100 flips. Y can take on values $0, 1, 2, \ldots 100$.

A rv is **continuous** if its set of possible values is an entire interval of numbers.

Examples: Let T be the time between two customers entering a store. Let U be a random number from the interval $[-5, 5]$.

Convention: We usually denote random variables by a capital letter near the end of the alphabet (e.g. X, Y) and specific instances of the random variable by a lower case letter.

Big Picture In statistics, we will model populations using random variables, and features/parameters (e.g. mean, variance) of these random variables will tell us about the population we are studying.

Probability mass function, pmf

Experiment: Roll a six-sided dice twice.
$S = \{(i,j) \mid i,j \in \{1,2,3,4,5,6\}\}$. Let X be the sum of the two rolls. X can take on values $2, 3, \ldots, 12$.

pmf - continued

Definition: A probability mass function of a discrete rv, X, is given by

$$p(x) = P(X = x) = P(\text{all } x \in S \mid X(s) = x)$$

What do we expect?

Example

A lab has 6 computers. Let X denote the number of these computers that are in use during the lunch hour. Suppose the pmf of X is given by:

x	0	1	2	3	4	5	6
p(x)	.05	.10	.15	.25	.20	.15	.1

Probability that at most 2 computers are in use:

Probability that at least half of the computers are in use

Probability that there are 3 or 4 computers free

Cumulative distribution function, cdf

The **cumulative distribution function (cdf)** is given by

$$F(y) = P(X \leq y) = \sum_{x \leq y} P(X = x)$$

x	0	1	2	3	4	5	6
p(x)	.05	.10	.15	.25	.20	.15	.1