Introduction to Random Variables

DASC 512

Overview

- Random Variables
- Discrete Random Variables
- Expectation
- Variance
- Continuous Random Variables
- Linear Combinations of Random Variables
- Other Probability Functions

Random Variables

- A random variable (RV) is a mapping that assigns numerical values to the possible outcomes of an experiment such that each sample point is represented by a unique numerical value.

If there are n possible discrete outcomes, $\{x_1, x_2, \dots, x_n\}$, then the random variable X maps those values to $i = 1,2, \dots, n$

The outcomes might not be numerical (e.g., colors), but the RV maps them to numerical values.

Discrete Random Variables

- A discrete random variable is a random variable that assumes a countable number of values

- The function that maps outcomes to probabilities is called the probability mass function (pmf)

Probability Mass Function

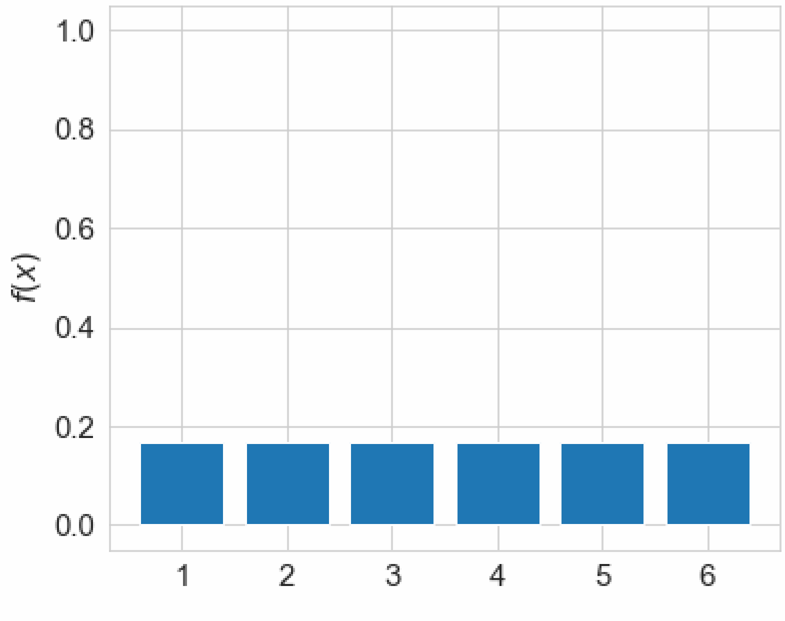
- To define a discrete random variable X
  - Define all possible outcomes (i.e., the sample space)
  - Assign a probability associated with each outcome

- The probability mass function (pmf) fully defines a distribution.

$$f(x) = P(X = x)$$

- Example: For a six-sided die:

$$f(x_i) = \frac{1}{6}, \qquad i = 1, 2, 3, 4, 5, 6$$

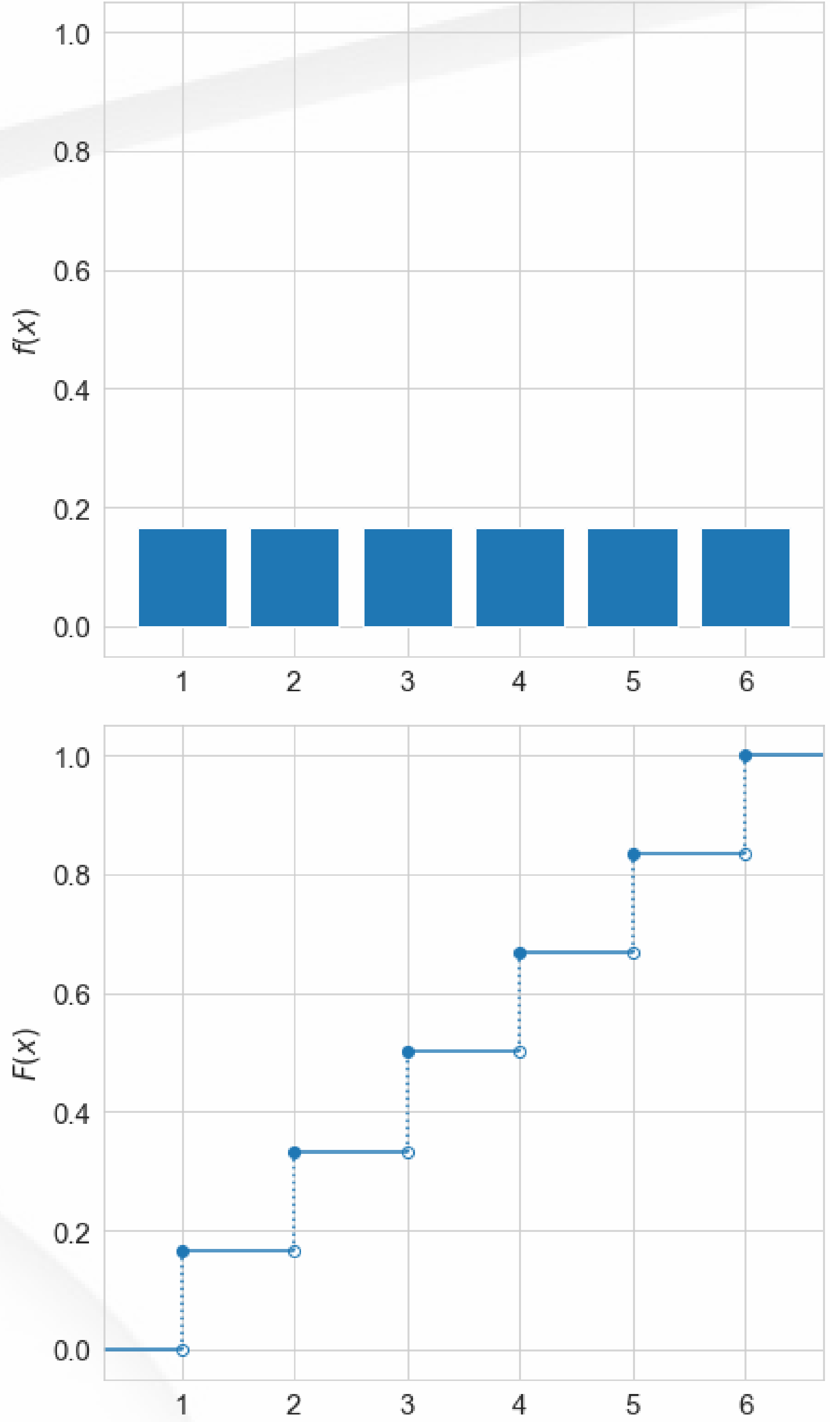


Cumulative Distribution Function

- The cumulative distribution function (cdf) also fully defines a distribution.

- The pmf f is
$$f(x) = P(X = x)$$

- The cdf F is the partial sum
$$F(x) = P(X \leq x)$$



Expectation

The expected value or expectation $E(X) = \mu$ is the mean observation expected to be observed from a random variable

$$\bar{x} \xrightarrow[n \to \infty]{} E(X) = \mu$$

It is calculated from the pmf

$$E(X) = \sum_{i=1}^{n} x_i f(x_i) = \sum_{i=1}^{n} x_i P(X = x_i)$$

Expectation

Example: For a 4-sided die,

$$E(X) = \frac{1}{4}(1) + \frac{1}{4}(2) + \frac{1}{4}(3) + \frac{1}{4}(4) = \frac{5}{2}$$

Variance

The <u>variance</u> ($\mathrm{Var}(\mathrm{X}) = \sigma^2$) is the expected squared deviation from the mean

$$s^2 \xrightarrow[n\to\infty]{} E\left((X-\mu)^2\right) = \sigma^2$$

It is also calculated from the pmf

$$Var(X) = E\left((X-\mu)^2\right) = \sum_{i=1}^{n}(x_i-\mu)^2 f(x_i) = \sum_{i=1}^{n}(x_i-\mu)^2 P(X=x_i)$$

Variance

Example: For a 4-sided die,

$$Var(X) = \frac{1}{4}\left(1 - \frac{5}{2}\right)^2 + \frac{1}{4}\left(2 - \frac{5}{2}\right)^2 + \frac{1}{4}\left(3 - \frac{5}{2}\right)^2 + \frac{1}{4}\left(4 - \frac{5}{2}\right)^2 = \frac{5}{4}$$

Standard Deviation

The standard deviation is the square root of the expected squared deviation from the mean

$$s \xrightarrow[n\to\infty]{} \sqrt{E\left((X-\mu)^2\right)} = \sqrt{\sigma^2} = \sigma$$

It is the square root of the variance.

Continuous Random Variables

- A continuous random variable is a random variable that assumes an uncountable number of values – values contained in one or more intervals

- The function that maps outcomes to probabilities is called the probability density function (pdf)

Probability Density Function

The area under the pdf $f(x)$ within an interval is the probability of observing an outcome from that interval.

$$\int_a^b f(x)\,dx = P(a \leq X \leq b) = P(a < X < b)$$

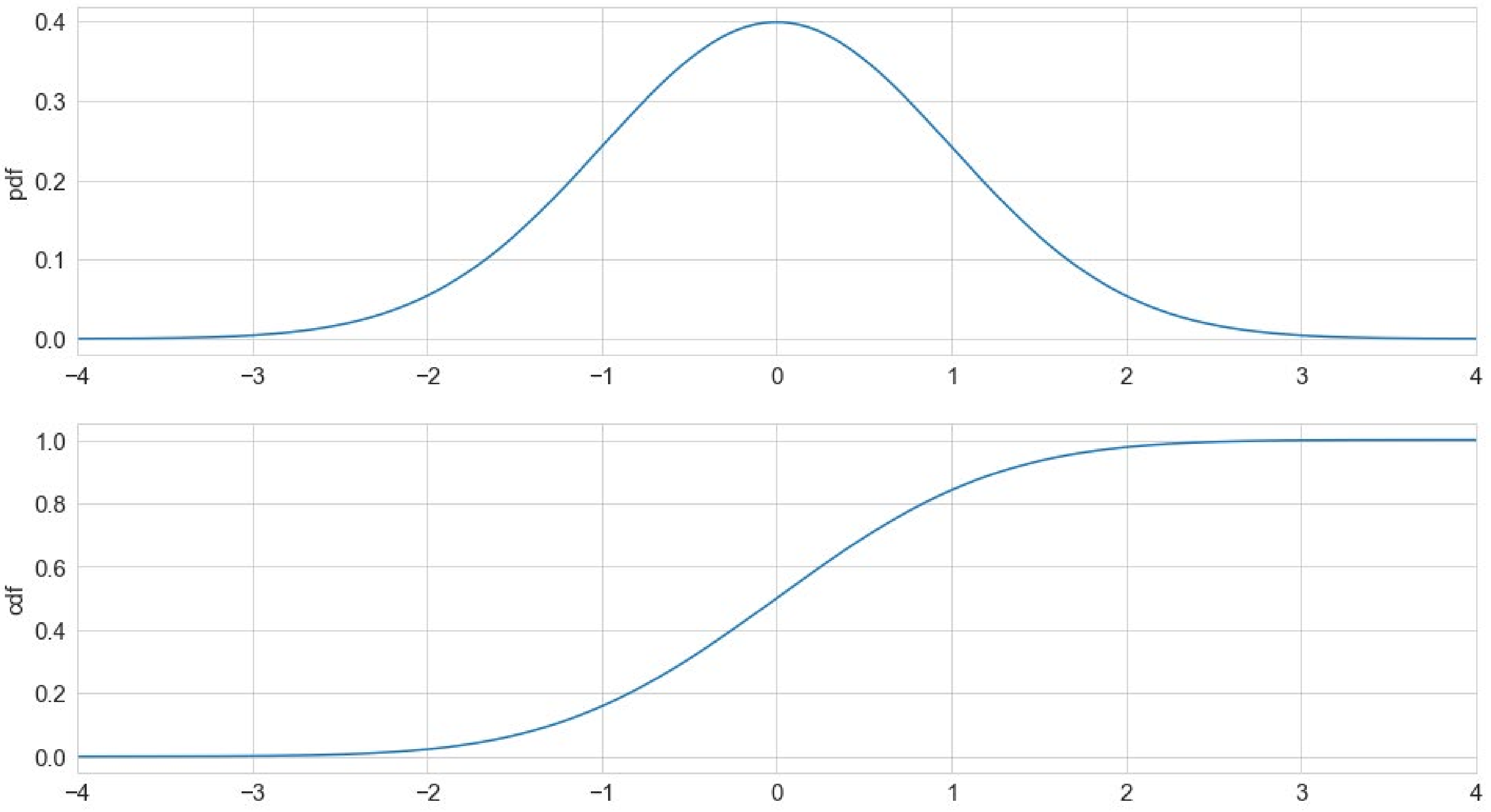
The probability of observing any specific value is 0.

Cumulative Distribution Function

- The cdf $F(x)$ is the area under the curve of the pdf to the left of x.
- As in the discrete case, it is the probability of observing a value smaller than x.

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$
$$\lim_{x \to -\infty} F(x) = 0$$
$$\lim_{x \to \infty} F(x) = 1$$

Example: Normal Distribution



Expectations of Continuous RVs

Calculation of E(X) and Var(X) for continuous RVs requires use of calculus

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx$$

Luckily, we can use existing solutions for named distributions

Linear Combinations

A linear combination of random variables X and Y can be written as

$$aX + bY$$

where a and b are any real numbers.

A linear combination of random variables is itself a random variable.

Linear Combinations

Any linear combination of random variables $aX + bY$ has expected value

$$E(aX + bY) = aE(X) + bE(Y) = a\mu_X + b\mu_Y$$

It also has variance

$$Var(aX + bY) = a^2Var(X) + b^2Var(Y) = \mathrm{a}^2\sigma_X^2 + b^2\sigma_Y^2$$

And by extension, it has standard deviation

$$\sigma_{aX+bY} = \sqrt{a^2\sigma_X^2 + b^2\sigma_Y^2}$$

Other Probability Functions

The pdf and cdf are commonly used, but Python includes several related functions that will be useful. We'll look at:

- Percentile Point Function (ppf)
  - Also known as quantile function or probit function (normal distribution only)
- Survival Function (sf)
- Inverse Survival Function (isf)

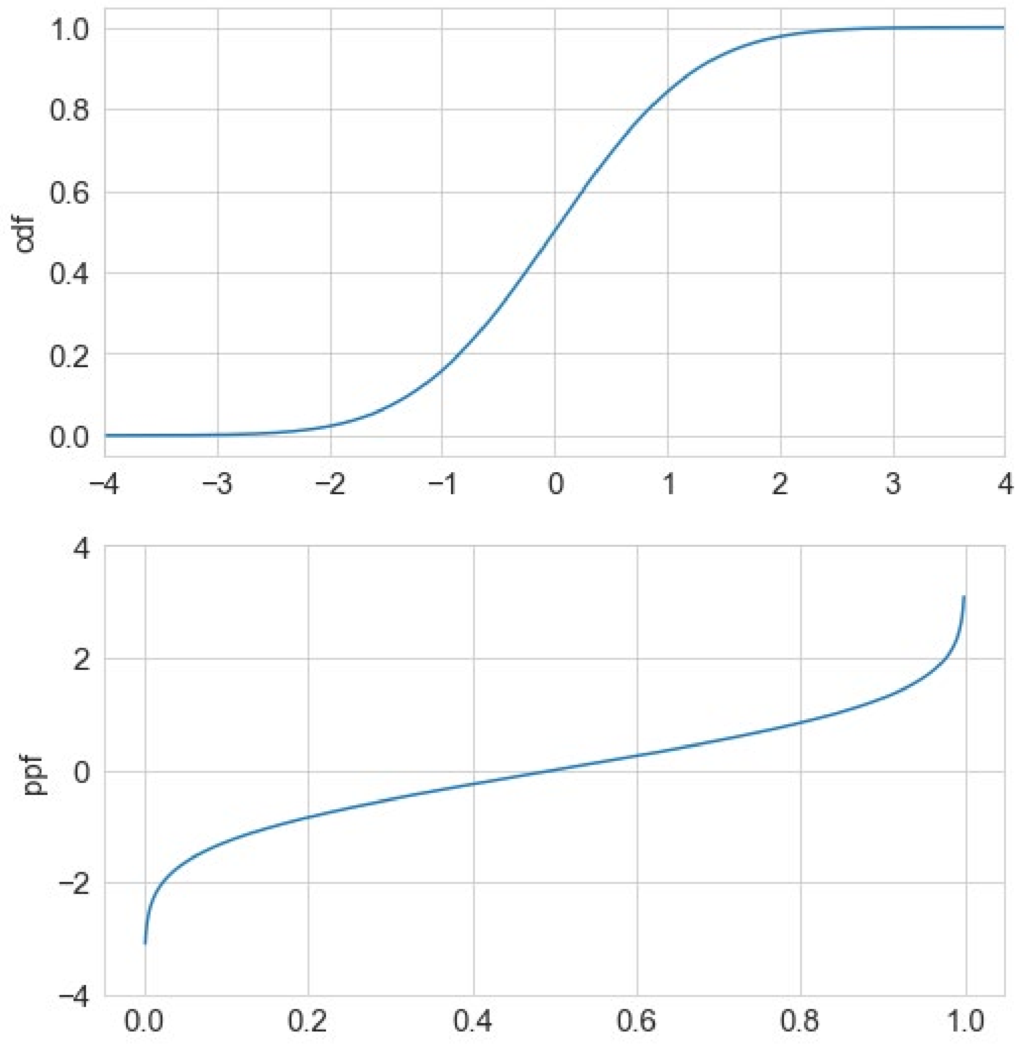
Percentile Point Function

The ppf is the inverse of the cdf

cdf: $x \rightarrow P(X \leq x)$

ppf: $P(X \leq x) \rightarrow x$

In other words, the input of the ppf is the desired quantile, and the ppf gives you that quantile

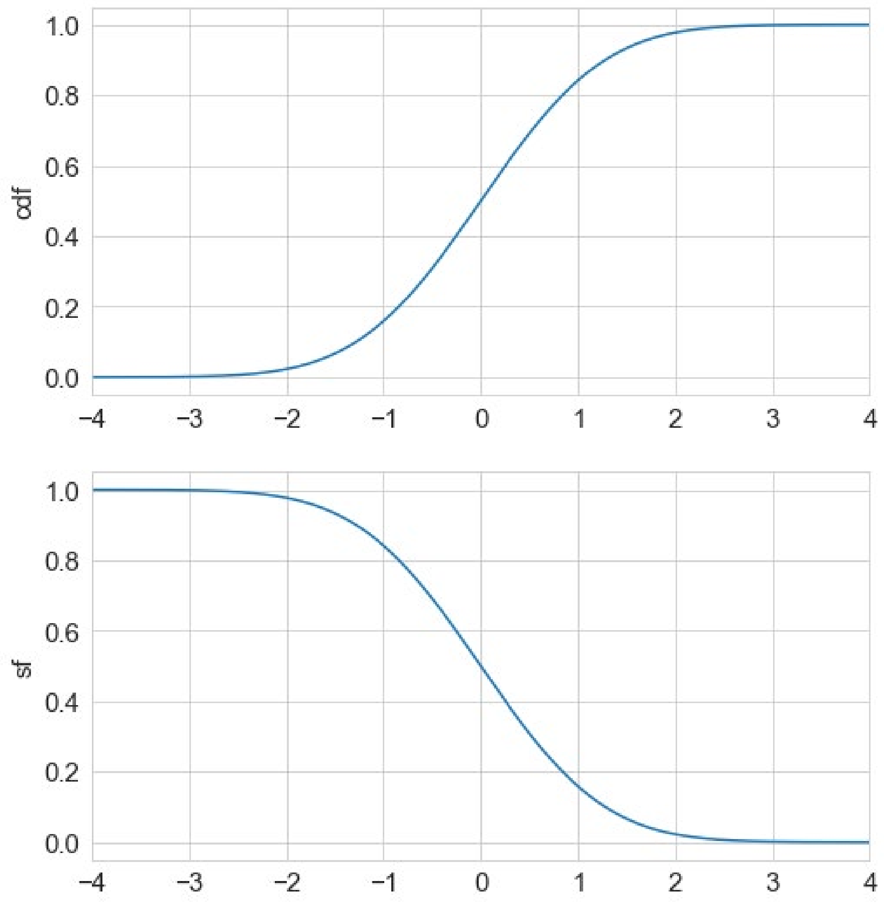


Survival Function

The sf is (1-cdf)

cdf: $x \rightarrow P(X \leq x)$

sf: $x \rightarrow P(X > x)$



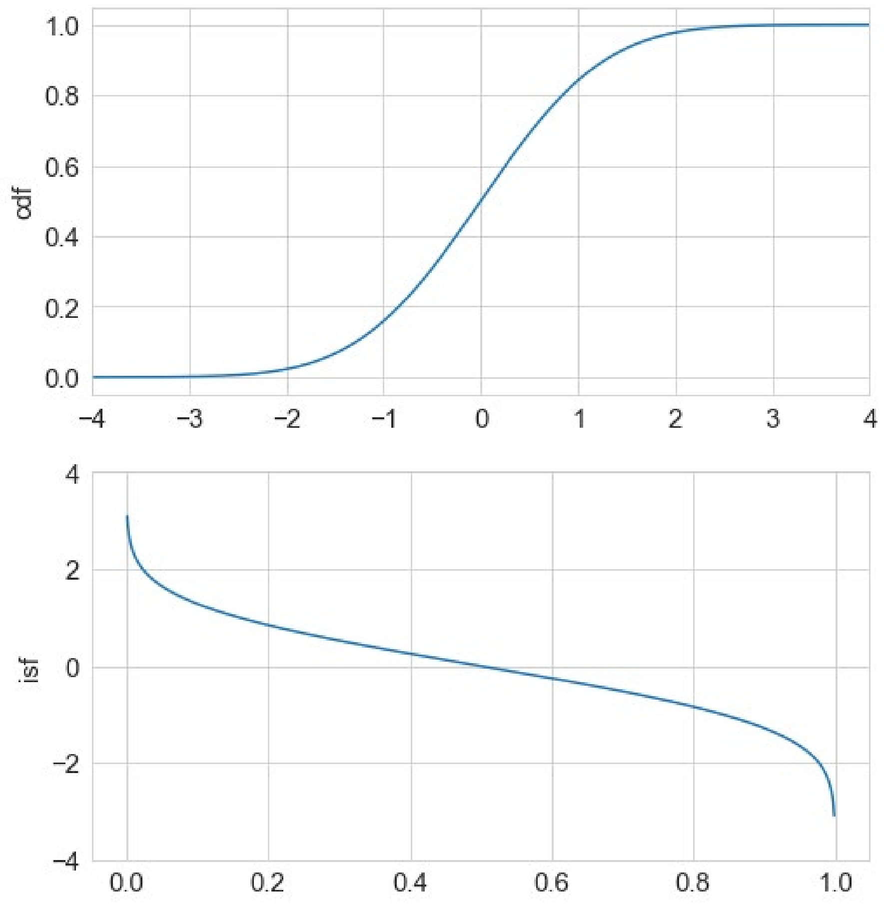
Inverse Survival Function

The isf is the inverse of the sf

cdf: $x \rightarrow P(X \leq x)$

sf: $x \rightarrow P(X > x)$

isf: $P(X > x) \rightarrow x$



Inverse Functions for Discrete Distributions

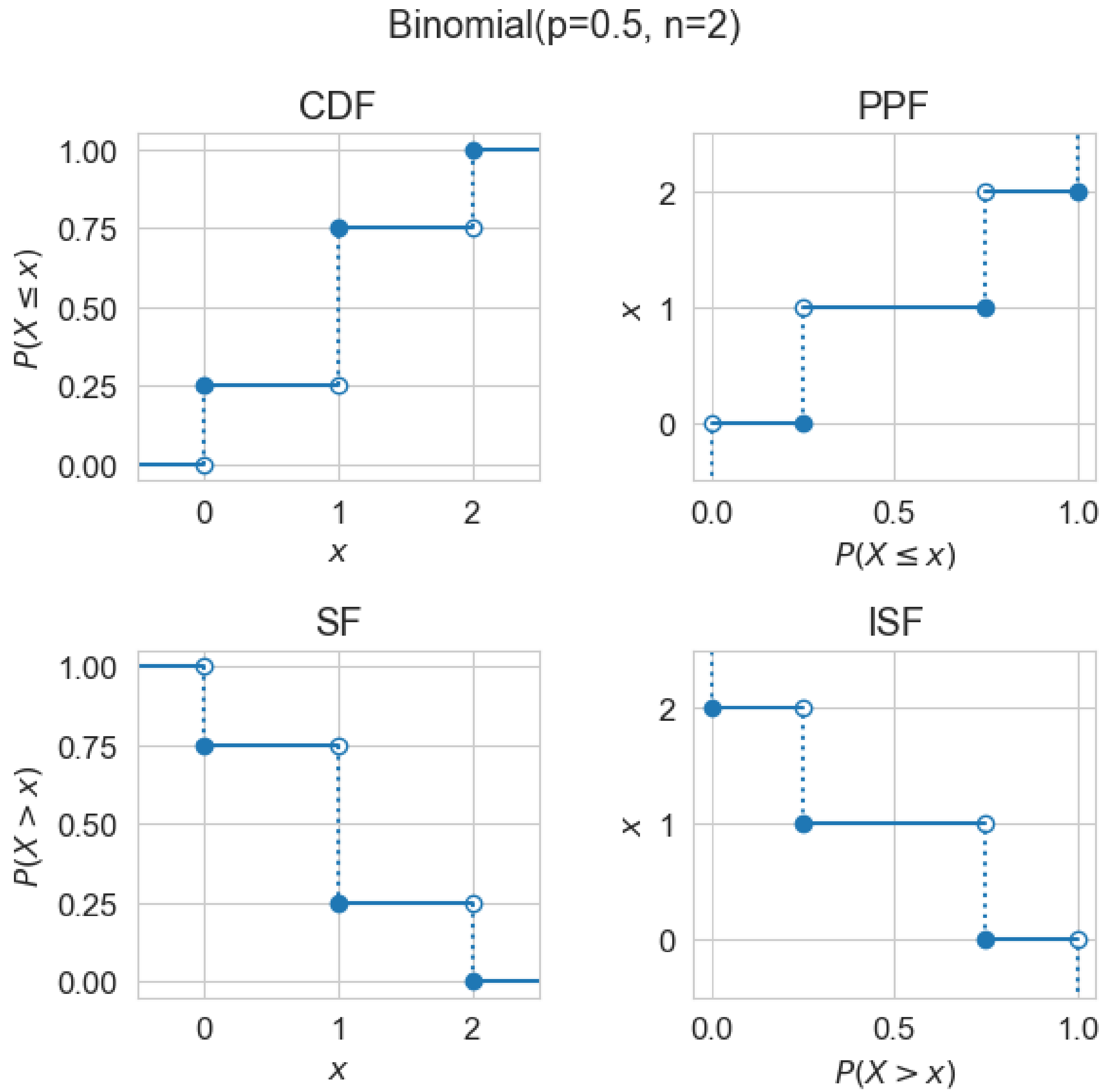
Interpretation is slightly different with discrete distributions

See at right, PPF maps onto dotted lines (e.g., F(x) is never 0.1)

For input quantile q,

PPF: $P(X \leq x) \geq q$

ISF: $1 - P(X \leq x) \leq q$



Summary of Probability Functions

First step to most problems in this class: define what you are solving for in terms of a probability

Function	**Input/Argument**	**Output**
pmf (discrete only)	x	$P(X = x)$
pdf (continuous only)	x	$f(x)$
cdf	x	$P(X \leq x)$
ppf	$P(X \leq x)$	x
sf	x	$P(X > x)$
isf	$P(X > x)$	x

Recap

- Random Variables
- Discrete Random Variables
- Expectation
- Variance
- Continuous Random Variables
- Linear Combinations of Random Variables
- Other Probability Functions