week01

August 17, 2021

1 Matrix Algebra

Statisticians and data scientists use the language of vectors and matrices to organize data and work with models. We will take a close look at matrix algebra, taking time to relate matrix algebra to algebra that you have worked with in the past, and show how to apply the tools in matrix algera to statistics you learned in Population Health Data Science I.

1.1 Vectors and operations with them

1.1.1 Definition of a vector

For this class we will consider a **vector** an ordered list of real numbers. To define a vector with the values 1,2,3,4,5, in that order, we can write

$$v = [1, 2, 3, 4, 5] \tag{1}$$

or

$$v = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \\ 5 \end{bmatrix} \tag{2}$$

A vector has a **length**, defined as the number of values included in that vector. For example, the above vector is length 5.

Vectors of length 2 and 3 can be visualized as directed line segments. To visualize a vector $v = [x, y]$ of length 2, place your pencil at the origin [0,0] and draw a line to the point $[x, y]$.

1.1.2 Vector times a scalar

A vector of length one is called a **scalar**. We can define the results of multiplying a vector $v = [a, b, c, d]$ by a scalar α as each entry of the vector times the scalar α.

$$\alpha v = \begin{bmatrix} \alpha \times a \\ \alpha \times b \\ \alpha \times c \\ \alpha \times d \end{bmatrix} \tag{3}$$

When a vector is multiplied by a scalar, that vector is "stretched" if the scalar is greater than one, "shrunk" if the scalar is less than one, and not changed if the scalar is equal to one.

1.1.3 Vector plus/minus a vector

A vector v plus a vector q creates a new vector that adds the individual entries of v and q

$$v = [4, -1, 7] \tag{4}$$
$$q = [0, 32, 9] \tag{5}$$
$$v + q = [4 + 0, -1 + 32, 7 + 9] = [4, 31, 16] \tag{6}$$

A vector v minus a vector q creates a new vector that subtracts the individual entries of q from v

$$v = [4, -1, 7] \tag{7}$$
$$q = [0, 32, 9] \tag{8}$$
$$v - q = [4 - 0, -1 - 32, 7 - 9] = [4, -33, -2] \tag{9}$$
$$q - v = [0 - 4, 32 - (-1), 9 - 7] = [-4, 33, 2] \tag{10}$$

1.1.4 Visualizing vector addition and subtraction

To visualize the result of adding and subtarcting two vectors, we can draw a parallellogram with the lengths of the sides of the parallellogram correspond to the legnths of the two vevtors we wish to add or subtract.

Suppose we want to add and subtract the vectors a and b

$$a = \begin{bmatrix} 1 \\ 2 \end{bmatrix} \tag{11}$$

$$b = \begin{bmatrix} 2 \\ 5 \end{bmatrix} \tag{12}$$

1.2 The Matrix

A **matrix** is a ordered list of vectors.

$$M = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 3 & 2 & 13 \\ 90 & 23 & 0 \end{bmatrix} \tag{13}$$

The **dimension** of a matrix is a an ordered pair of numbers: the first number indicates the number of rows of the matrix and the second number indicates the number of columns. For example, M has dimension

$$(4, 3)$$

.

1.2.1 Matrix Transpose

Given a matrix M, the **tranpose** of M, $M^{'}$ or M^{T}, is a matrix where the first row of $M^{'}$ corresponds to the first column of M, the second row of $M^{'}$ corresponds to the second columns of M and so on. If M has dimension (r, c) than $M^{'}$ has dimensions (c, r).

For example, if

$$M = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 3 & 2 & 13 \\ 90 & 23 & 0 \end{bmatrix} \tag{14}$$

then the tranpose of M,

$$M^{'} = \begin{bmatrix} 1 & 4 & 3 & 90 \\ 2 & 5 & 2 & 23 \\ 3 & 6 & 13 & 0 \end{bmatrix} \tag{15}$$

1.2.2 Matrix plus/minus a matrix

Given a matrix A and matrix B, the sum of A and B is a new matrix C where the i,j entry of C (C_{ij}) is the sum of the corresponding entries from A and B ($C_{ij} = A_{ij} + B_{ij}$). To add two matrices they must have the same dimension.

$$A = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \end{bmatrix} \tag{16}$$

$$B = \begin{bmatrix} 6 & 5 & 4 \\ 3 & 2 & 1 \end{bmatrix} \tag{17}$$

$$C = A + B = \begin{bmatrix} 1+6 & 2+5 & 3+4 \\ 4+3 & 5+2 & 6+1 \end{bmatrix} = \begin{bmatrix} 7 & 7 & 7 \\ 7 & 7 & 7 \end{bmatrix} \tag{18}$$

$$D = A - B = \begin{bmatrix} 1-6 & 2-5 & 3-4 \\ 4-3 & 5-2 & 6-1 \end{bmatrix} = \begin{bmatrix} -5 & -3 & -1 \\ 1 & 3 & 5 \end{bmatrix} \tag{19}$$

$$\tag{20}$$

1.2.3 Matrix times a Matrix

Two matrices A and B can be multiplied together $C = AB$ if the number of columns of A is equal to the number of rows of B. The i,j entry of this product, C_{ij}, assuming th number of columns of A and number of rows of B equals N, is the following sum of products:

$$C_{i,j} = a_{i,1}b_{1,j} + a_{i,2}b_{2,j} + a_{i,3}b_{3,j} + \cdots + a_{i,N}b_{N,j} \tag{21}$$

$$= \sum_{e=1}^{N} a_{i,e}b_{e,j} \tag{22}$$

For example, suppose that

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} \tag{23}$$

and that

$$B = \begin{bmatrix} 1 & 2 & -1 \\ 3 & 4 & -2 \end{bmatrix}. \tag{24}$$

Then the product $C = AB$ equals

$$C = AB = \begin{bmatrix} 1*1+2*3 & 1*2+2*4 & 1*-1+2*-2 \\ 3*1+4*3 & 3*2+4*4 & 3*-1+4*-2 \end{bmatrix} = \begin{bmatrix} 7 & 10 & -5 \\ 15 & 22 & -11 \end{bmatrix} \tag{25}$$

it is important to note that AB is not necessarily equal to BA.

1.2.4 Matrix times a vector

A matrix A with dimension (r, c) can be multiplied by a vector v if the length of v is c. Then the product Av is a vector with the the i$^{\text{th}}$ entry of Av defined as

$$(Av)_i = \sum_{k=1}^{c} A_{i,k} v_k \tag{26}$$

For example, Let the matrix

$$A = \begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \end{bmatrix} \tag{27}$$

and the vector

$$v = \begin{bmatrix} 4 \\ -3 \end{bmatrix} \tag{28}$$

then

$$Av = \begin{bmatrix} 1*4+2*-3 \\ 3*4+4*-3 \\ 5*4+6*-3 \end{bmatrix} = \begin{bmatrix} -2 \\ 0 \\ 2 \end{bmatrix} \tag{29}$$

1.2.5 Matrix inverse

The identity matrix The **identity matrix** of dimension r, usually labled I_r, is a matrix with r rows and r columns such that the diagonal elements of the matrix, the (1,1) entry, (2,2) entry, up to (r,r) entry are the number 1 and all other entries are 0. For example,

$$I_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{30}$$

This matrix is called the identity matrix because any matrix A times I returns A (try it!). To be more precise, this matrix is the *multiplicative* identity matrix. But the adjective *multiplicative* is usually dropped.

The inverse matrix The **inverse** of a matrix A—A^{-1}—is a matrix such that, if A has the same number of rows and columns (called square) then

$$AA^{'} = A^{'}A = I \tag{31}$$

1.2.6 The matrix as a function

A function, by definition, is a mathematical object that takes an input and returns a unique output. For example the function $f(x) = x^2$ defined on the real numbers takes as input any real number and returns one, unique real number. The function $f(x, y) = xy$ takes as input a pair (x, y) and returns the product of x and y. One input corresponds to one output.

From above, when we multiply a matrix by a vector the result is a single vector. For every vector v, Av returns a vector. We can think of A as a function. We input the vector v to matrix A and receive an output vector (Av).

1.2.7 The dot (inner) product and orthogonality

The **dot product** between vector v with entries $[v_1, v_2, \cdots, v_l]$ and vector q with entries $[q_1, q_2, \cdots, q_l]$ is defined as

$$v \cdot q = v'q = v_1q_1 + v_2q_2 + \cdots, v_lq_l \tag{32}$$

$$= \sum_{i=1}^{l} v_iq_i \tag{33}$$

We can use the dot product to help define matrix multiplication. For a matrix A and B, the (i,j) entry of the product $C = AB$ is

$$C_{i,j} = \sum_{k=1}^{N} a_{i,k}b_{k,j} \tag{34}$$

$$= a_{i,}^{'}b_{,j} \tag{35}$$

where $a_{i,}$ denotes the i^{th} row of the matrix A and $b_{,j}$ denotes the j^{th} column of the matrix B.

1.2.8 The norm

For this class, every vector we encounter has a **norm**. The norm for a vector v with entries $v_1, v_2, v_3, \cdots, v_N$ will be defined in this class as

$$||v|| = \left(v_1^2 + v_2^2 + v_3^2 + \cdots + v_N^2\right)^{1/2} \tag{36}$$

$$= \left(\sum_{i=1}^{N} v_i^2\right)^{1/2} \tag{37}$$

$$= v'v \tag{38}$$

In the case of vectors with length 2 this norm should look familiar

$$v = [a, b] \tag{39}$$

$$||v|| = \sqrt{a^2 + b^2} \tag{40}$$

1.2.9 Ways matrix algebra can help us think about staistics

Matrix algebra is frequently used in statistics because of how this algebra simplifies working with multiple data points. For example, assume we collected N obervations and for each observation we collected $p + 1$ pieces of information: the first p pieces of information are meant to help us explain the $p + 1$ piece of info that we will call y.

Our data can be formatted into a matrix with rows corresponding to observations and p columns, one column for each piece of information:

$$D = \begin{bmatrix} d_{1,1} & d_{1,2} & \cdots & d_{1,p} \\ d_{2,1} & d_{2,2} & \cdots & d_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N,1} & d_{N,2} & \cdots & d_{N,p} \end{bmatrix} \tag{41}$$

and y, our target or response variable we wish to explain with the above p variables, can be formamted as a vector

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_N \end{bmatrix} \tag{42}$$

With our data and target defined as above, we can use vectors to create famailar statistics. The mean of the p pieces of information we collected on N observations—often called p covariates—can be computed using a vectors of 1s

Define a ones vector of length N as

$$1_N = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix} \tag{43}$$

Then the mean of each covariate in D is

$$\frac{1_N' D}{N} = \left[\frac{\sum_{i=1}^N d_{i,1}}{N}, \frac{\sum_{i=1}^N d_{i,2}}{N}, \cdots, \frac{\sum_{i=1}^N d_{i,p}}{N} \right] \tag{44}$$

In simple linear regression, we related each target y_i to the sum of a linear function $\beta_0 + \beta_1 x_i$ and an error term (ϵ) that is Normally distributed.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon \tag{45}$$
$$\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right) \tag{46}$$

We can rewrite the above using matrix algebra. First, we find a dot product between two vectors hiding in plain sight

$$y_i = \beta_0 + \beta_1 x_i \tag{47}$$
$$y_i = \beta_0 1 + \beta_1 x_i \tag{48}$$
$$y_i = X_i \beta \tag{49}$$

where

$$X_i = \begin{bmatrix} 1 & x_i \end{bmatrix} \tag{50}$$
$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{51}$$

The above describes a single, i$^{\text{th}}$ observation

$$y_i = X_i \beta \tag{52}$$
$$\epsilon_i \sim \mathcal{N}\left(0, \sigma^2\right), \tag{53}$$

but we can describe all N observations at once without much chance to the above

$$y = X\beta \tag{54}$$
$$\epsilon \sim \mathcal{N}\left(0, \sigma^2\right), \tag{55}$$

where

$$X = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ \vdots & \\ 1 & x_N \end{bmatrix} \tag{56}$$
$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \tag{57}$$
$$\epsilon = \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_N \end{bmatrix} \tag{58}$$

1.3 Review of important distributions of random variables

[66]:
```
def pdfAndCDF(f,F,xmin,xmax,discrete=True):
    import numpy as np
    import scipy

    fig,axs = plt.subplots(1,2)

    if discrete:
        domain = np.arange(xmin,xmax+1)

        ax=axs[0]
        for x in domain:
            ax.plot([x]*2,[0,f(x)],color="steelblue")
            ax.scatter(x,f(x),color="steelblue")
        ax.set_xlabel("X values",fontsize=12)
        ax.set_ylabel("Probability mass (density) function",fontsize=12)
        ax.set_xticks(domain)

        ax=axs[1]
        for x in domain:
            ax.plot([x]*2,[0,F(x)],color="steelblue")
            ax.scatter(x,F(x),color="steelblue")
        ax.set_xlabel("X values",fontsize=12)
        ax.set_ylabel("Cumulative mass (density) function",fontsize=12)
        ax.set_xticks(domain)
    else:
        domain = np.linspace(xmin,xmax,10**3)

        ax=axs[0]
        ax.plot(domain,f(domain),color="steelblue")
        ax.set_xlabel("X values",fontsize=12)
        ax.set_ylabel("Probability mass (density) function",fontsize=12)

        ax=axs[1]
        ax.plot(domain,F(domain),color="steelblue")
        ax.set_xlabel("X values",fontsize=12)
        ax.set_ylabel("Cumulative mass (density) function",fontsize=12)

    fig.set_size_inches(8,4)
    fig.set_tight_layout(True)
    plt.show()
```

1.4 Discrete Distributions

1.4.1 Bernoulli

Definition A random variable X follows a **Bernoulli** distribution if X takes either the value 0 or 1, and

$$p(x) = \begin{cases} 0 & (1-\theta) \\ 1 & \theta \end{cases} \tag{59}$$

We write that $X \sim \text{Bern}(\theta)$, in words, that X follows a Bernoulli distribution with parameter theta.

Expected value and variance The expected value of X, if it is distributed Bernoulli is

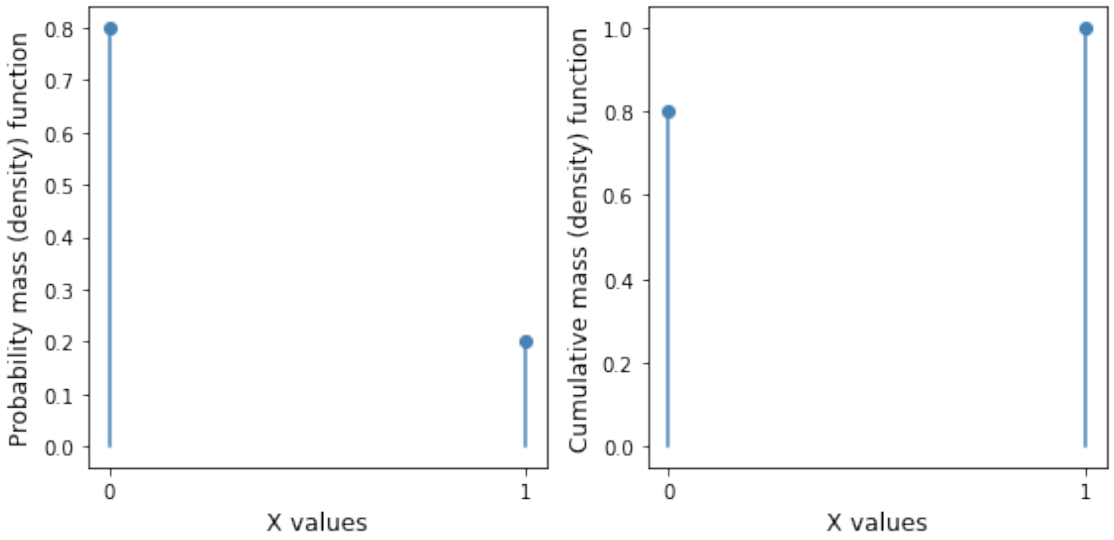
$$\begin{aligned} \text{E}(X) &= 1 \times p(X=1) + 0 \times p(X=0) && (60)\\ &= 1 \times \theta + 0 \times (1-\theta) && (61)\\ &= \theta && (62) \end{aligned}$$

The variance of X is

$$\begin{aligned} \text{Var}(X) &= (1-\theta)^2 \times p(X=1) + (0-\theta)^2 \times p(X=0) && (63)\\ &= (1-\theta)^2 \times \theta + \theta^2 \times (1-\theta) && (64)\\ &= \theta(1-\theta)\left[(1-\theta)+\theta\right] && (65)\\ &= \theta(1-\theta) && (66) \end{aligned}$$

Below is a plot of the pdf and cdf for a Bernoulli(0.2) random variable

[67]:
```
import scipy
pdfAndCDF(scipy.stats.binom(1,0.2).pmf, scipy.stats.binom(1,0.2).
 ↪cdf,xmin=0,xmax=1,discrete=True)
```



1.4.2 Geometric

Definition A random variable has a geometric distribution if it is defined on all non-negative integers and the probability distribution of these values is

$$p(X = t) = (1 - p)^{t-1} p \tag{67}$$

This type of random variable describes a sequence of trials (experiments, outcomes, etc) where the first $t - 1$ trials "failed" and the final trial was a "success". For example, the number of shots until you make a free-throw on shot t or the number of negative COVID-19 tests until the t^{th} test is positive could both have a geometric distribution.

The values of a random variable are the number of attempts until a success.

Expected value and variance If a r.v. X has a geometric distribution with parameter p,

$$X \sim \text{Geom}(p) \tag{68}$$

The expected value is

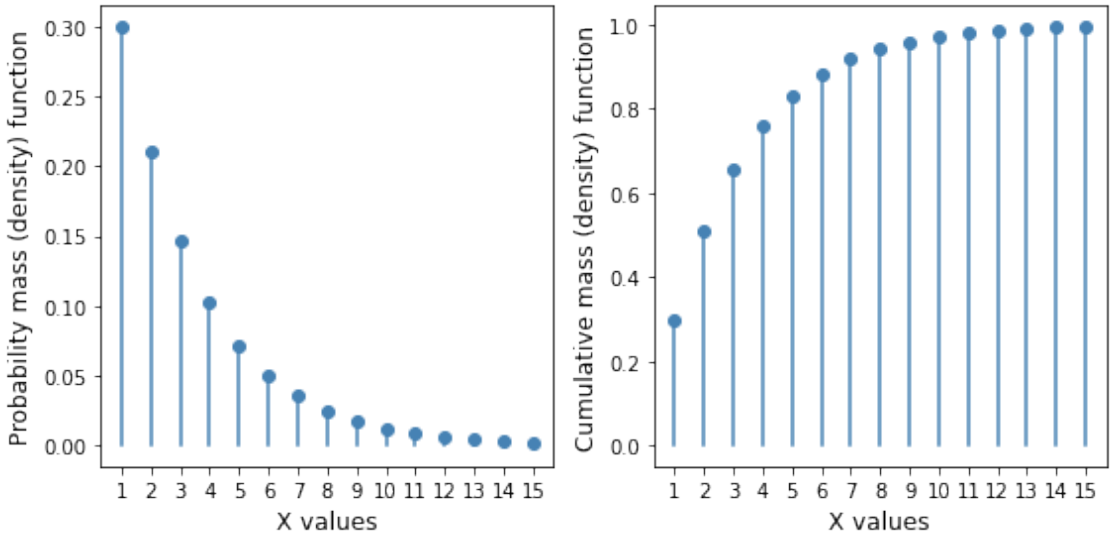
$$E(X) = \frac{1}{p} \tag{69}$$

and the variance is

$$\text{Var}(X) = \frac{1}{p} \frac{(1 - p)}{p} \tag{70}$$

Below is a plot of a Geometric(0.3) random variable for the values from 1 to 15.

[68]:
```
import scipy
pdfAndCDF(scipy.stats.geom(p=0.3).pmf,scipy.stats.geom(p=0.3).cdf,1,15)
```



1.4.3 Uniform (Discrete)

Definition A random variable has a uniform discrete distribution—$X \sim U(a, b)$—if it is defined for all values between two parameters a and b, and the probability of values at or between a and b is

$$p(x) = \frac{1}{N} \tag{71}$$

where N is the number of values the random variable can be. The uniform discrete distribution assigns an equal probability to all possible values of a random variable. #### Expected value and variance

The expectation of a r.v. following a uniform discrete distribution is the average of the two endpoints

$$E(X) = \frac{a+b}{2} \tag{72}$$

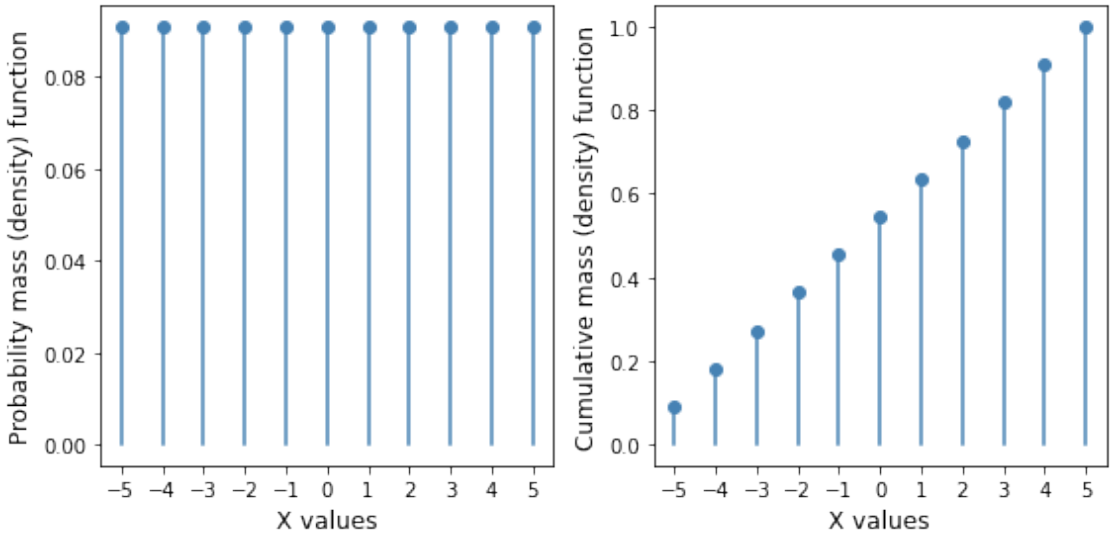
The variance is

$$Var(X) = \frac{(b-a+1)^2 - 1}{12} \tag{73}$$

Below is a plot of the pmf and cmf for a Uniform Discrete(-5,5) random variable.

[69]:
```
import scipy
```

```
pdfAndCDF(scipy.stats.randint(-5,5+1).pmf,scipy.stats.randint(-5,5+1).
↪cdf,-5,5,discrete=True)
```



1.4.4 Binomial

Definition The binomial distribution is a discrete distribution assigning probabilities to a number of successes (usually assigned a value of 1) out of a total number of N. We assume each of the N trials are independent from one another, that only a success or failure can occur, and that a success occurs with a probability θ. The binomial distribution is used anytime you ask "what is the probability of x occurrences of the same event out of a total N number of tries?"

The probability mass function (discrete so we can assign probabilities to individual outcomes) is

$$p(x) = \binom{N}{x}\theta^x(1-\theta)^{N-x} \tag{74}$$

The symbol $\binom{N}{x}$ is called the a binomial coefficient, sometimes said "N choose x". The binomial coefficient (we'll see below) is the number of times you can select x items from a total of N items without caring about the order you selected them.

Expected value and variance The expectation is

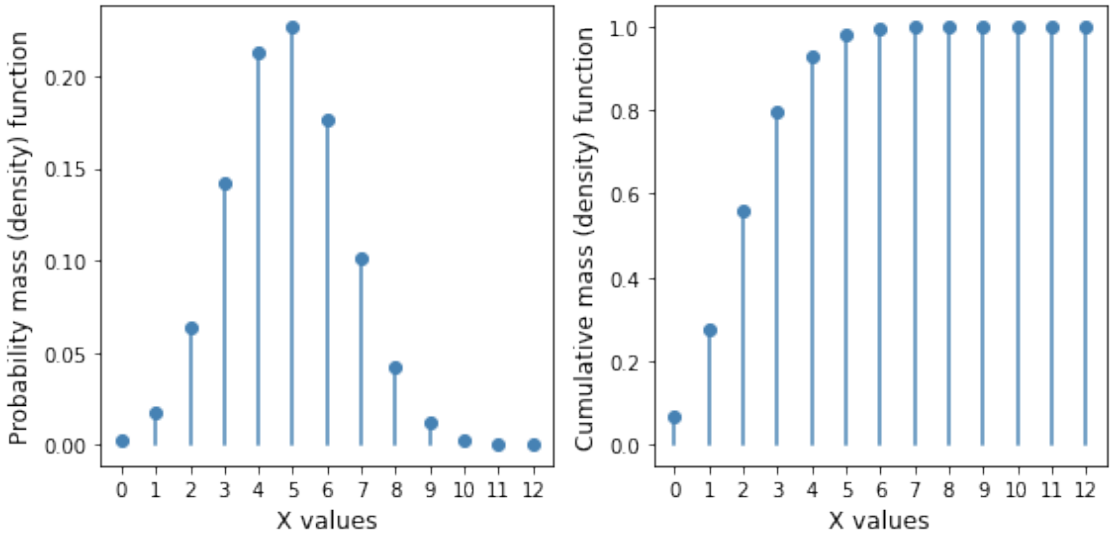
$$E(x) = N\theta \tag{75}$$

and variance is

$$Var(x) = N\theta(1-\theta) \tag{76}$$

Below is a plot of the pmf and cmf of a Binomial(12,0.4) distributed random variable.

[70]:
```
import scipy
pdfAndCDF(scipy.stats.binom(12,0.4).pmf,scipy.stats.binom(12,0.2).
 ↪cdf,0,12,discrete=True)
```



1.4.5 Poisson

Definition A random variable X has a Poisson distribution—$X \sim \text{Pois}(\lambda)$— if it is discrete and it's probability mass function is

$$p(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} \tag{77}$$

A Poisson distributed random variable assigns probabilities to the number of events that occur over a specific time period. The parameter λ is typically thought of as the rate of events per unit of time. For example, the number of deaths per month, number of phone calls per week, or number of emails per hour (so many).

1.4.6 Expected value and variance

The expected value is

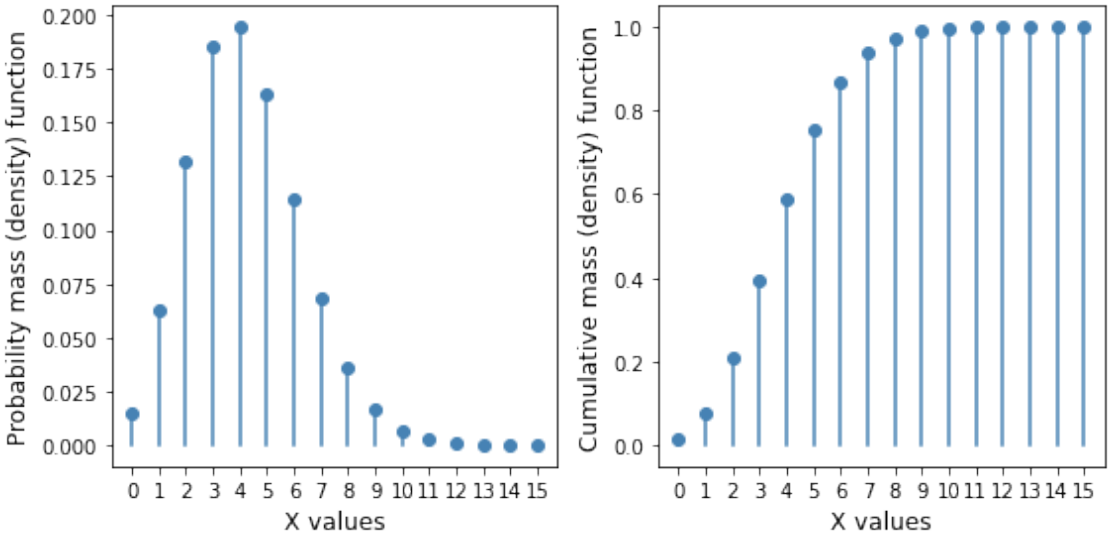
$$\mathrm{E}(X) = \lambda \tag{78}$$

and the variance is

$$\text{Var}(X) = \lambda \tag{79}$$

Below is the pmf and cdf for a random variable with a Poisson(4.2) distribution for the values from 0 to 15.

[71]:
```
import scipy
pdfAndCDF(scipy.stats.poisson(4.2).pmf,scipy.stats.poisson(4.2).
 ↪cdf,0,15,discrete=True)
```



1.5 Continuous Distributions

Continuous probability distributions, unlike discrete distributions, do not assign probabilities to every possible outcome. This is because the probability of any single value a continuous r.v. could take is zero.

We associate probabilities with a **probability density function** (pdf), defining probability on a continuous *interval* as the areas under the pdf. When our r.v. was discrete, the sum of the probabilities of all possible events had to equal one. There is a similar rule for continuous r.v.s and densities. The sum of the area under the curve of a pdf associated with a r.v. X over the largest possible interval must equal 1.

Like the cumulative mass function for discrete r.v.s, continuous random variables have a **cumulative density function (CDF)**. The CDF is a function that assigns a *probability* from the smallest possible value of a random variable up to a user-specified input (say x). Because the input to a *CDF* is an interval—(smallest possible value, x)—the output is a probability.

1.5.1 Uniform (continuous)

A r.v. has a continuous uniform distribution $X \sim U(a, b)$ if the probability density is

$$f(x) = \frac{1}{b-a} \tag{80}$$

The probability is defined for intervals between a and b. The CDF for a continuous uniform distribution is

$$F(x) = \frac{x}{b-a} \tag{81}$$

Expectation and variance The expected value (expectation) equals

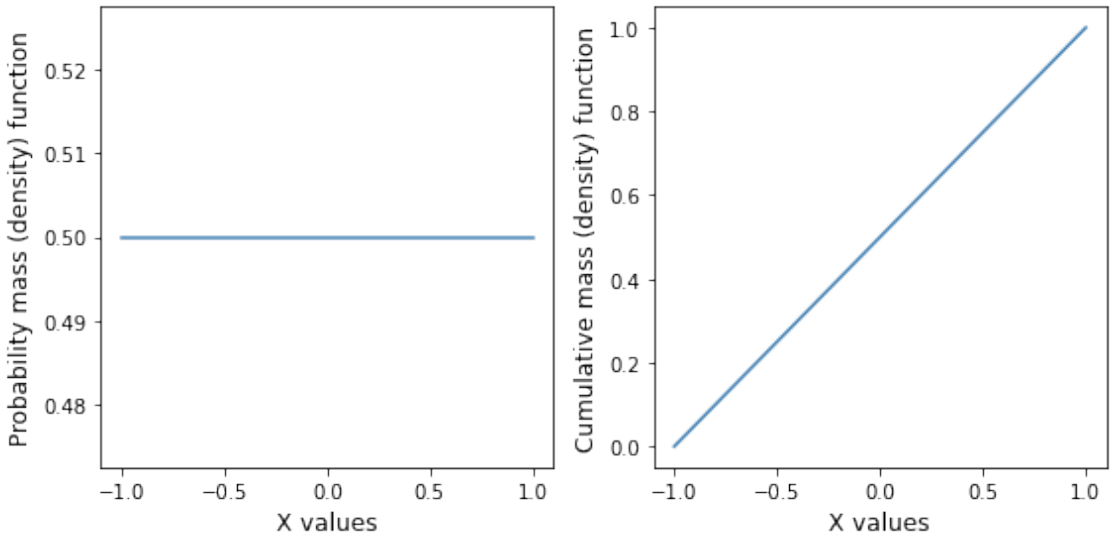
$$E(X) = \frac{a+b}{2} \tag{82}$$

and the variance equals

$$Var(X) = \frac{(b-a)^2}{12} \tag{83}$$

Below is the pmf and cdf for a random variable with a Uniform(-1,1) distribution.

[72]:
```
import scipy
pdfAndCDF(scipy.stats.uniform(-1,2).pdf,scipy.stats.uniform(-1,2).
 ↪cdf,-1,1,discrete=False)
```



1.5.2 Normal (Gaussian)

The Normal (or Gaussian) distribution describes the probability of a continuous random variable. This is the (likely familiar) bell curve.

The Normal distribution has two parameters: the mean (μ) and the standard deviation (σ). The pdf is symmetric and unimodel and defined over all values from negative infinity to positive infinity. Values close to the mean are much, much more likely that values further from the mean. Because of this, the Normal distribution describes phenomena or values of a r.v. that are more or less expected to be close to μ—surprises are not very likely.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma}} \tag{84}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma}\right\} \tag{85}$$

$$= \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \tag{86}$$

Expectation and variance If X is a r.v. and normally distributed $X \sim \mathcal{N}\left(\mu, \sigma^2\right)$, the expectation is

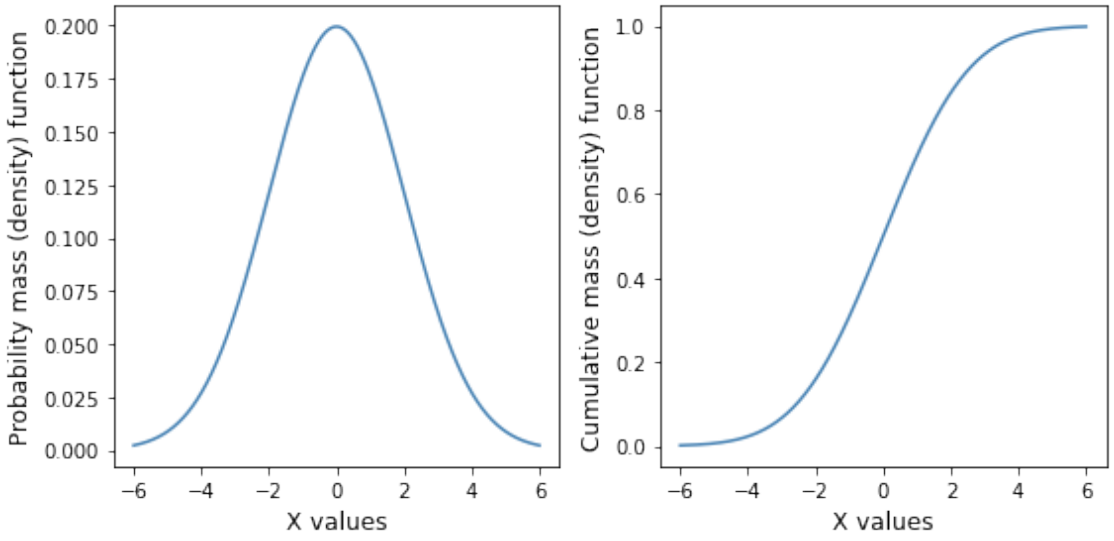
$$E(x) = \mu \tag{87}$$

and variance is

$$Var(x) = \sigma^2 \tag{88}$$

Below is the pmf and cdf for a random variable with a Normal(0,2) distribution.

[73]:
```
import scipy
pdfAndCDF(scipy.stats.norm(0,2).pdf,scipy.stats.norm(0,2).
 ↪cdf,-6,6,discrete=False)
```



1.5.3 Multivariate normal distribution

The multivariate normal distribution describes the probability over vectors of length N such that each entry of the vector alone has a normal distribution.

Definition A vector of length n has a multivariate normal distribution (mvn)

$$x \sim \text{MVN}(\mu, \Sigma) \tag{89}$$

where μ is a vector of length n and Σ is a matrix of dimension (n, n) if the probability density associated with x equals

$$f(x) = (2\pi)^{-n/2} |\Sigma|^{-1/2} exp\left\{\frac{1}{2}(x-\mu)'\Sigma^{-1}(x-\mu)\right\} \tag{90}$$

Expectation and variance The expected value of x and variance of x are

$$E(x) = \mu \tag{91}$$
$$Var(x) = \Sigma \tag{92}$$

The matrix Σ is called the covariance matrix. This matrix is square (same number of rows and columns) and symmetric (the element Σ_{ij} = the element Σ_{ji}). The covariance matrix has on it's diagonals the individual variances for each element x. And the off diagonal elements, Σ_{ij} contains the covariance between x_i and x_j.

1.6 The expected value

1.6.1 Expectation for discrete random variables

The **expectation** of a random variable is the sum of each value of the random variable weighted by the probability that value will occur.

If X is a random variable then the expectation of X is

$$\mathrm{E}(X) = x_1 p(x_1) + x_2 p(x_2) + \cdots + x_n p(x_n) \tag{93}$$

$$\sum_{i=1}^{N} x_i p(x_i) \tag{94}$$

where the random variable X can take any of n values from x_1 to x_n.

1.6.2 The Expectation of a function of a r.v.

We can generalize the expectation of a random $[E(X)]$ allowing us to compute the expectation of any function of a random variable. The expected value of a function of a r.v. X is

$$E[f(X)] = \sum_{i=1}^{N} f(x_i) p(x_i) \tag{95}$$

The expected value of $f(X)$ is just the value of each $f(x_i)$ weighted by how often the value x_i occurs. To see that this expectation is more general than our original definition, we can compute $E[f(X)]$ where f is the identity function ($f(x) = x$).

$$E[f(X)] = \sum_{i=1}^{N} f(x_i) p(x_i) \tag{96}$$

$$= \sum_{i=1}^{N} x_i p(x_i) \tag{97}$$

$$= E(X) \tag{98}$$

1.7 The Law of Large Numbers

The Law of Large Numbers says, informally, that a random variable's sample mean approaches its expected value as the number of observations used to compute the sample mean increases. The mathematical statement is more precise:

$$\text{For any } \epsilon > 0 \text{ there exists a } N \text{ such that } p(|\bar{X}_n - E(X)| < \epsilon) = 1 \tag{99}$$

where $\bar{X}_n$ is the sample mean of N independent random variables $X_1, X_2, \cdots, X_n$ with the same expected value ($E(X)$).