week01

August 11, 2021

0.1 Stat Recap

For week01 we'll spend time reacquainting ourselves with the fundementals of probability and statistics, and learn how to use a few new, valuable statistical tools.

0.2 Probability

0.3 Universal Set, Outcomes, and Events

We can think of **probability** as a function that assigns a number from 0 to one (inclusive) to items in a set. The set of all items is called the **Universal Set** or **Sample Space**. Items in the Sample space are called **outcomes** and defined as results of an experiment or stochatic, random occurance. **Events** are any set of outcomes. Two events are **disjoint** is they have no outcomes in common, or $A \cap B = \varnothing$ where $\cap$ is set intersection and $\$\varnothing\$istheemptyset, thesetofnoitems$.

0.3.1 Example 01

Let us suppose our experiment is to roll a single six-sided die and record the number that appears on the top face. The potential *outcomes* are the numbers 1, 2, 3, 4, 5, and 6. Our *sample space* is the set $\{1, 2, 3, 4, 5, 6\}$. There are six outcomes: the individual numbers 1,2,3,4,5,6. An example of some events are $\{1, 3, 5\}$ or odd numbers, $\{2, 4, 6\}$ or even numbers, and the event $\{5\}$.

0.3.2 Example 02

0.3.3 Example 03

0.4 Set theoretic definition of probability and the Kolmorgorov Axioms

0.4.1 A probability definition

With the notion of a sample space, outcomes, and an event we can define the probability of an event as the number of items in the event divided by the number of items in the sample space.

$$p(E) = \frac{\text{Number of items in E}}{\text{Number of items in the Sample space}} \tag{1}$$

0.4.2 Kolmogorov Axioms

Let S be the Sample space,E an event, and let A and B be disjoint events. The **Kolmorogov Axioms** formalize our intuition about probability: - KA1: $p(S) = 1$ (Something has to happen) - KA2: $0 \leq p(E) \leq 1$ - KA3: $p(A \cup B)$ = p(A) + p(B) where $\cup$ is set union.

0.4.3 Nice results due to KA1-3

Nice result 01 The Sample space (S) can also be characterized by the items in a set A and it's complement $A^{'}$—the items not in the set A—$S = A \cup A^{'}$. But the sets A and $A^{'}$ are disjoint and so

$$p(S) = p(A \cup A^{'}) = p(A) + p(A^{'}) = 1 \tag{2}$$
$$p(A) = 1 - p(A^{'}) \tag{3}$$

Nice result 02 If the event A is a subset of B—all the items in A are also items that belong to B—than intuitively we expect the probability of B to be the same or bigger than the probability of A. KA1-3 help us show that. We can decompose the set B into the items that belong to A and the items that belong to B that are not in A (lets call this set C), and then use the fact that they are disjoint.

$$p(B) = p(A \cup C) = p(A) + p(C) \geq p(A) \tag{4}$$

Nice result 03 The complement of the Sample space, the set of all outcomes, is the empty set, and so

$$1 = p(S) = p(S \cup \varnothing) = p(S) + p(\varnothing) \tag{5}$$
$$1 = p(S) + p(\varnothing) \tag{6}$$
$$1 = 1 + p(\varnothing) \tag{7}$$
$$0 = p(\varnothing) \tag{8}$$

0.5 Conditional Probability, Bayes Theorem, and the Law of Total Probability

0.5.1 Cond. Prob

A conditional probability is the probability of an event given another event has already occured. If we are interested in the probability of an event B given A has occured we write $P(B|A)$. The conditional probability of B given A is defined as

$$p(B|A) = \frac{p(B \cap A)}{p(A)} \tag{9}$$

This probability can be thought of as counting the number of outcomes in B that are also outcomes in A and dividing by the sample space restricted to the set A. The **multiplication rule** of probability comes in handy often and is a rearrangement of the definition of conditional probability.

$$p(A \cap B) = p(B|A)p(A) \tag{10}$$

0.5.2 Bayes Theorem

Bayes Theorem is derived from the multiplicatin rule and allows us one way to relate two conditional probabilities to one another.

$$p(A \cap B) = p(B|A)p(A) \tag{11}$$
$$p(B \cap A) = p(A|B)p(B) \tag{12}$$
$$\tag{13}$$

and because $p(A \cap B) = p(B \cap A)$ we find

$$p(B|A)p(A) = p(A|B)p(B) \tag{14}$$
$$p(B|A) = \frac{p(A|B)p(B)}{p(A)} = p(A|B)\frac{p(B)}{p(A)} \tag{15}$$

0.5.3 Law of total probability

The law of total probability allows us to break the probability of a single event (A) into the sum of several smaller events that are related to A. Suppose we know that there exist sets $B_1, B_2, \cdots, B_N$ that cover all possible outcomes in A and we also know that the intersection of any B_i with A is disjoint.

Then

$$p(A) = \sum_{i=1}^{N} p(A \cap B_i) = \sum_{i=1}^{N} p(A|B_i)p(B_i) \tag{16}$$

0.6 Transition from Probability to Statistics

0.7 Random variables

A **random variable** is a *function* that maps outcomes in the sample space to real numbers (positive, negative, and decimal numbers). A random variable is not by itself random. Instead, a random variable depends on the results of an experiment which is random.

0.7.1 Example 01

Suppose we roll a six-sided die and record the number on the top side of the die. We can create many diffeent random variables from this experiment. For example, we can create the random variable X that maps event numbers to the value one and odd numbers to the value 0:

$$X = \begin{cases} 1 \text{ when the outcome is even} \\ 0 \text{ when the outcome is odd} \end{cases} \tag{17}$$

We can create a second random variable Y that maps the number on the top of the die to the corresponding integer.

$$\begin{aligned} Y(1) &= 1 \\ Y(2) &= 2 \\ Y(3) &= 3 \\ Y(4) &= 4 \\ Y(5) &= 5 \\ Y(6) &= 6 \end{aligned} \tag{18}$$

0.7.2 How to compute probabilities for random variables

The random variables X and Y are not random in and of themselves, they are deterministic functions. What makes X and Y random is the unknown outcome from rolling a die. Because random variables take on value from the real number line with chance, it makes sense to talk about statements like the "the probability X equal one" or "the probability Y is greater than the value 3".

The most straight forward approach to computing the probability that a random variable equals some value v is to count all possible outcomes that are mapped by that random variable to v and divided by all outcomes in the sample space.

$$p(X = v) = \frac{|\{w \text{ such that } X(w) = v\}|}{\text{Number of items in the sample space}} \tag{19}$$

where $||$ around a set counts the number of items inside, and is called a set's **cardinality**.

0.7.3 The probability and cumulative mass function

A **probability mass function (f)** associated with the random variable X is a function that maps each value of X to it's corresponding probability.

$$f(x) \rightarrow \text{ the probability that } X \text{ equals x} \tag{20}$$

In other words,

$$f(x) = p(X = x) \tag{21}$$

for all possible values x.

A **cumulative mass function** associated with the random variable X is a function that maps each value of X to the probability the random variable is *less than that value*.

$$F(x) \rightarrow \text{ the probability that } X \text{ is less than or equal to x} \tag{22}$$

In other words,

$$F(x) = p(X \leq x) \tag{23}$$

for all possible values x.

[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()
```

0.8 Discrete Distributions

0.8.1 Bernoulli

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

$$p(x) = \begin{cases} 0 & (1-\theta) \\ 1 & \theta \end{cases} \tag{24}$$

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

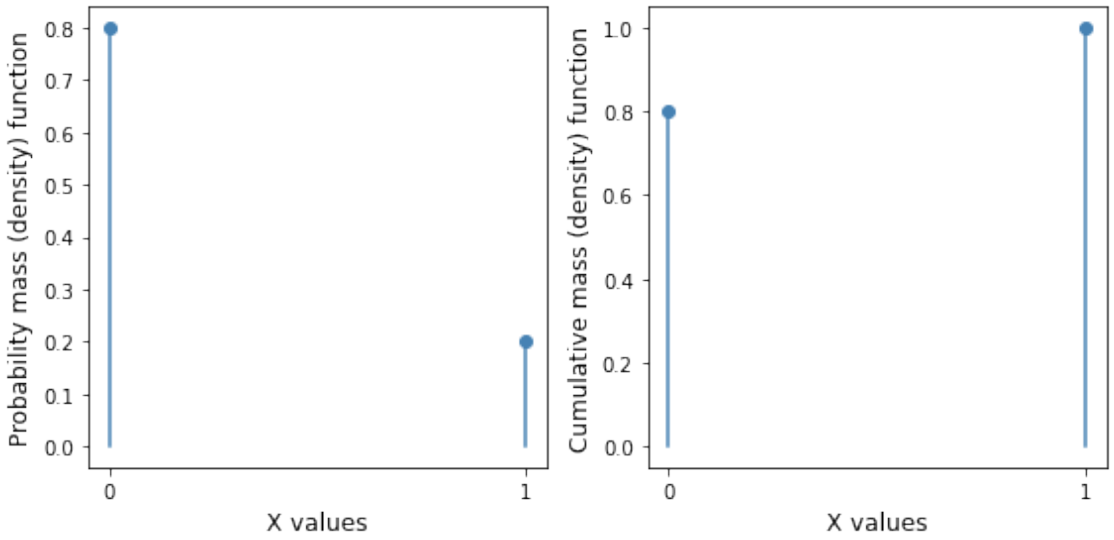
$$\begin{aligned} \text{E}(X) &= 1 \times p(X=1) + 0 \times p(X=0) && (25)\\ &= 1 \times \theta + 0 \times (1-\theta) && (26)\\ &= \theta && (27) \end{aligned}$$

The variance of X is

$$\begin{aligned} \text{Var}(X) &= (1-\theta)^2 \times p(X=1) + (0-\theta)^2 \times p(X=0) && (28)\\ &= (1-\theta)^2 \times \theta + \theta^2 \times (1-\theta) && (29)\\ &= \theta(1-\theta)\left[(1-\theta)+\theta\right] && (30)\\ &= \theta(1-\theta) && (31) \end{aligned}$$

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)
```



0.8.2 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

$$p(X = t) = (1 - p)^{t-1} p \tag{32}$$

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,

$$X \sim \text{Geom}(p) \tag{33}$$

The expected value is

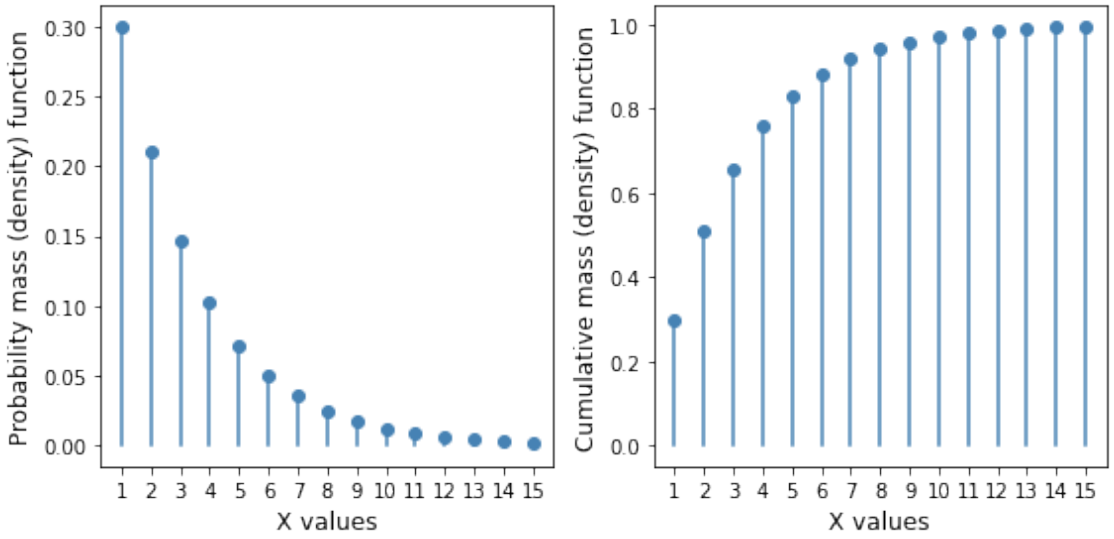
$$E(X) = \frac{1}{p} \tag{34}$$

and the variance is

$$\text{Var}(X) = \frac{1}{p} \frac{(1 - p)}{p} \tag{35}$$

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)
```



0.8.3 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

$$p(x) = \frac{1}{N} \tag{36}$$

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

$$E(X) = \frac{a+b}{2} \tag{37}$$

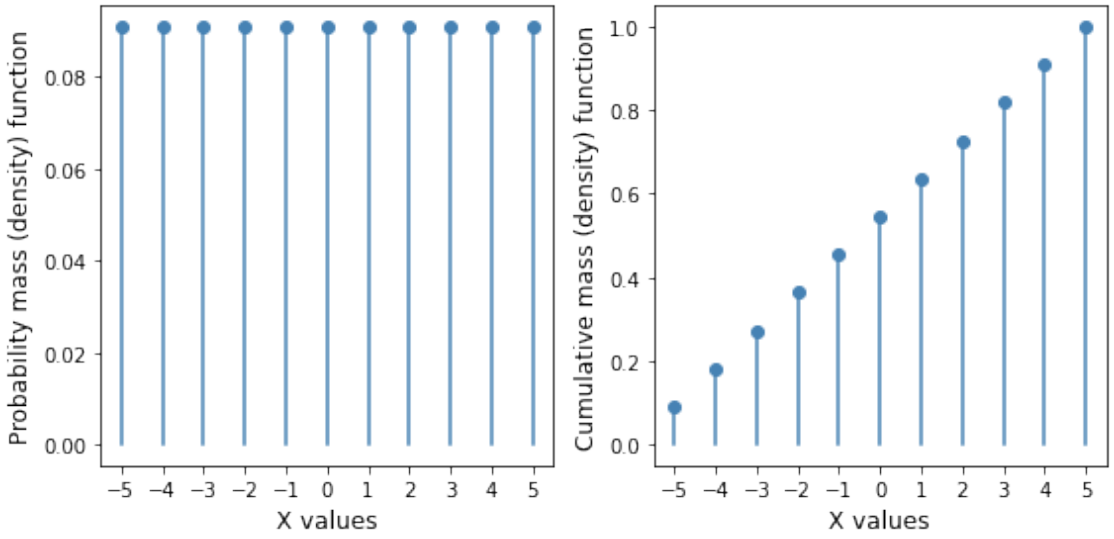
The variance is

$$Var(X) = \frac{(b-a+1)^2 - 1}{12} \tag{38}$$

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)
```



0.8.4 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

$$p(x) = \binom{N}{x}\theta^x(1-\theta)^{N-x} \tag{39}$$

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

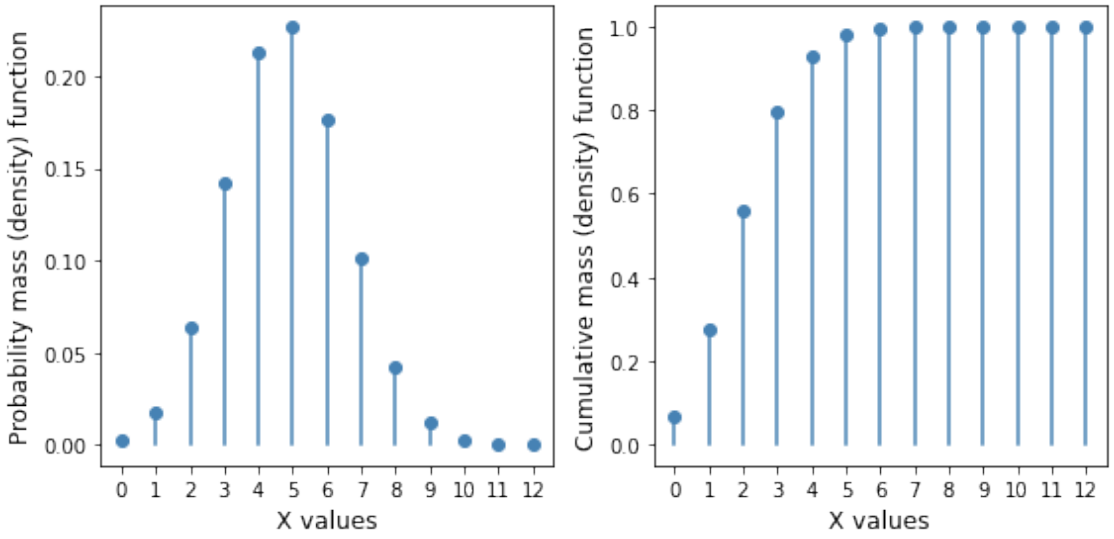
$$E(x) = N\theta \tag{40}$$

and variance is

$$Var(x) = N\theta(1-\theta) \tag{41}$$

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)
```



0.8.5 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

$$p(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} \tag{42}$$

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).

0.8.6 Expected value and variance

The expected value is

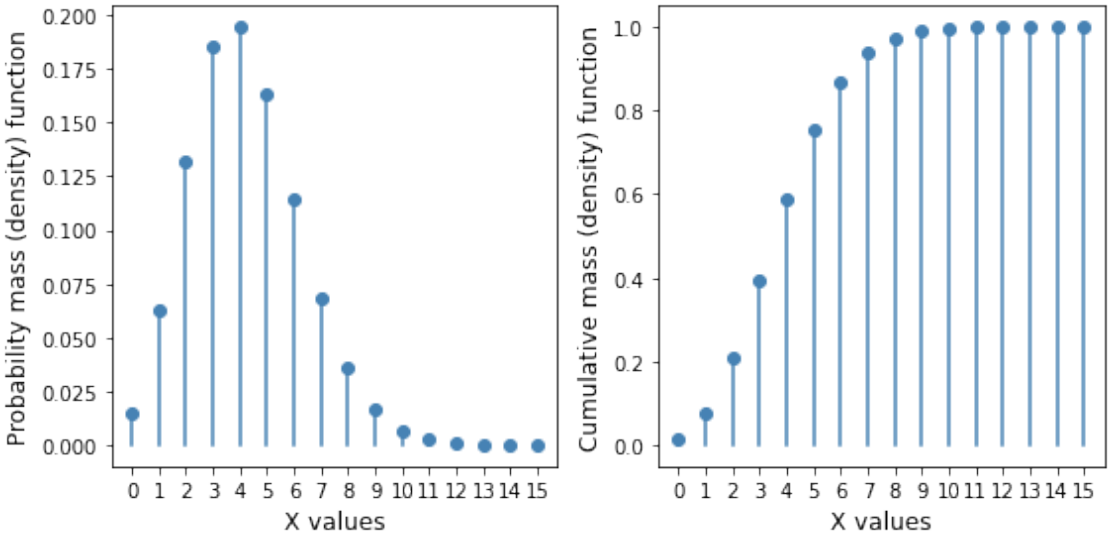
$$\mathrm{E}(X) = \lambda \tag{43}$$

and the variance is

$$\text{Var}(X) = \lambda \tag{44}$$

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)
```



0.9 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.

0.9.1 Uniform (continuous)

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

$$f(x) = \frac{1}{b-a} \tag{45}$$

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

$$F(x) = \frac{x}{b-a} \tag{46}$$

Expectation and variance The expected value (expectation) equals

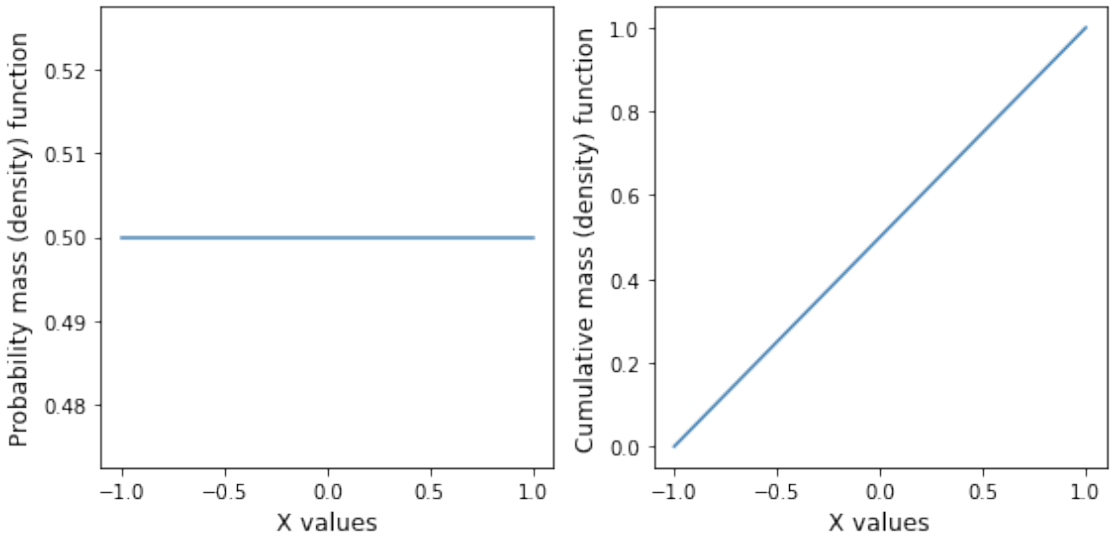
$$E(X) = \frac{a+b}{2} \tag{47}$$

and the variance equals

$$Var(X) = \frac{(b-a)^2}{12} \tag{48}$$

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)
```



0.9.2 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.

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma}} \tag{49}$$

$$= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma}\right\} \tag{50}$$

$$= \left(2\pi\sigma^2\right)^{-1/2} \exp\left\{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right\} \tag{51}$$

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

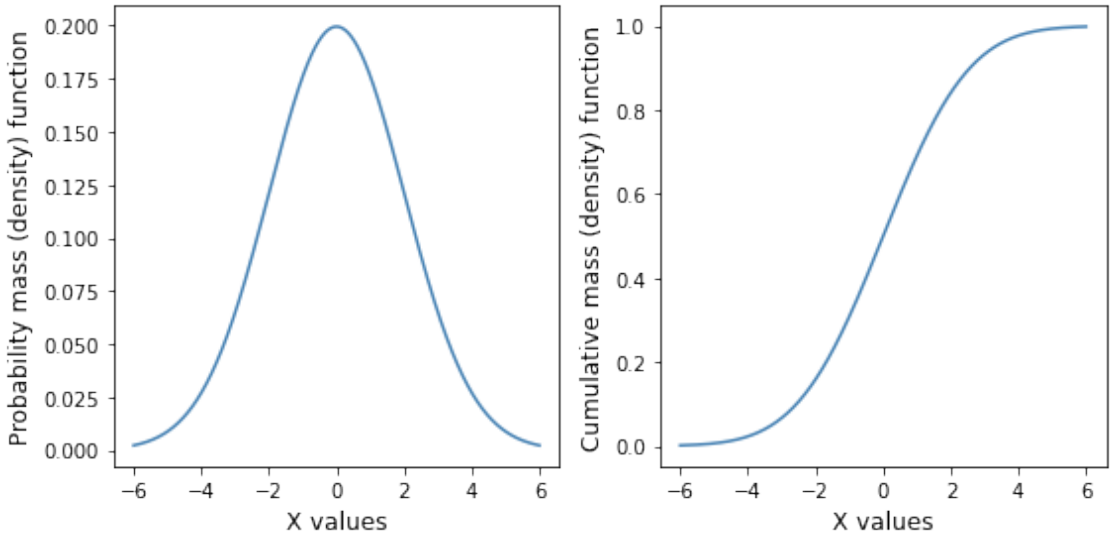
$$E(x) = \mu \tag{52}$$

and variance is

$$Var(x) = \sigma^2 \tag{53}$$

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)
```



0.10 The expected value

0.10.1 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

$$\mathrm{E}(X) = x_1 p(x_1) + x_2 p(x_2) + \cdots + x_n p(x_n) \tag{54}$$

$$\sum_{i=1}^{N} x_i p(x_i) \tag{55}$$

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

0.10.2 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

$$E[f(X)] = \sum_{i=1}^{N} f(x_i) p(x_i) \tag{56}$$

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 ($f(x) = x$).

$$E[f(X)] = \sum_{i=1}^{N} f(x_i)p(x_i) \tag{57}$$

$$= \sum_{i=1}^{N} x_i p(x_i) \tag{58}$$

$$= E(X) \tag{59}$$

0.11 The Central limit Theorem

0.11.1 Central limit theorem (informal description)

The distribution of the mean of successive samples from a population is well approximated by a normal distribution. But any normal distribution is defined by two parameters μ and σ. Given a sample of data $\mathcal{D}$, the first parameter (μ) can be approximated by the mean of the dataset $\bar{\mathcal{D}}$ and the second parameter (σ) by the **standard error** of the data.

0.11.2 Standard error

If $\mathcal{D} = [d_1, d_2, \cdots, d_n]$ is a dataset with values d_1,d_2 and so on, the standard error is defined as

$$\text{SE}_{\mathcal{D}} = \frac{s_{\mathcal{D}}}{\sqrt{n}} \tag{60}$$

where $s_{\mathcal{D}}$ is the standard deviation and n is the number of data points.

So then the central limit theorem says (roughly) that for a r.v. X with **any** distribution,

$$\bar{X} \sim \mathcal{N}\left(\bar{x}, \frac{s_{\mathcal{D}}}{\sqrt{n}}\right) \tag{61}$$

where $\bar{x}$ is the mean of values from X, $s_{\mathcal{D}}$ is the standard deviation, and n is the number of data points.

0.11.3 Z scores and "standardizing"

The normal distribution also plays a role in "standardizing" data. We can standardize a variable X with the following algorithm: 1. Compute the mean of X 2. Compute the standard deviation of X 3. For each value in X 1. subtract the mean 2. divide the value in 3.1. by the standard deviation

By standardizing, the values of X are put in terms of units of standard deviation. A value of 0 from sample means that sample value is the same as the population mean of the data. A value of 1 from a sample means that sample value is one standard deviation larger than the population mean and so on.

Standardizing is important when you want to compare two variables that have different units, or that have different variances around their mean.

When we are given a random variable X and create a new random variable $Z = \frac{X - \bar{X}}{sd(X)}$ by subtracting the mean of X and dividing by the standard deviation that new random variable is called a zscore.

0.12 Confidence intervals

0.12.1 Definition

A confidence interval quantifies the range of possible values a sample statistic can be within. Typically, a confidence interval is centered around the mean and is of the form

$$(\bar{x} - L, \bar{x} + L) \tag{62}$$

Confidence intervals have associated with them (not surprisingly given the name) a specific confidence, a value between 0 and 1 often turned into a percentage like "a 95% confidence interval".

With a Y% confidence interval we can make a statement like: If we sampled our data over and over and constructed a Y% confidence interval then %Y of those confidence intervals should contain the population parameter of interest.

0.12.2 CI for the mean (Applying the CLT)

We can use the Central limit theorem (CLT) to build a Y% confidence interval around the mean of a r.v. The CLT says that the sampling distribution of the mean is a normal distribution centered on the mean of our data (μ) and σ equal to the standard error. And so a reasonable way to construct a confidence interval for the true μ is

$$\text{Confidence interval} = (\bar{x} - z * \text{SE}_x, \bar{x} + z * \text{SE}_x) \tag{63}$$

But what is this z in our confidence interval equation for the mean?

A z-score (z) is the number of standard deviations a away from the mean of a random variable Z. The z-score is associated with a Normal distribution and every z-score is associated with a probability as follows:

If Z is a Normally distributed r.v. then

$$p(Z < \text{z-score}) = \text{the area under the normal curve up until z} \tag{64}$$

0.12.3 Two-sided confidence interval

To compute a Y% confidence interval, you need to find the corresponding z-score (z) so that the area under the curve from $-z$ to $+z$ is $Y/100$. For example, the area under the curve between -1.96 and 1.96 is 0.95.

This then corresponds to a 95% confidence interval of a r.v. X as follows

We can create different % confidence intervals for the mean of a random variable by using different z scores. The probability between a z-score of -1.28 and 1.28 is 0.80 and so an 80% confidence interval for the mean of a random variable is

0.13 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:

$$\text{For any } \epsilon > 0 \text{ there exists a } N \text{ such that } p(|\bar{X}_n - E(X)| < \epsilon) = 1 \tag{65}$$

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 ($E(X)$).