# Assignment1

February 1, 2024

## 1 Assignment 1: Solutions

All problems have an equal weight of 10pts each. Each assignment is equally weighted.

[2]:

```
import numpy as np
import matplotlib.pyplot as plt
import math
```

### 1.1 1. Infectious Disease Testing

**A new infectious disease has emerged, and a diagnostic test has been developed to identify individuals who are infected. The test is not perfect and can yield both false positives and false negatives. You are given the following information:**

- **The prevalence of the disease in the population is 5%, meaning that 5% of individuals are actually infected.**
- **The sensitivity of the test is 90%, which means it correctly identifies 90% of infected individuals as positive.**
- **The specificity of the test is 85%, which means it correctly identifies 85% of uninfected individuals as negative.**

**Part A. [4pts] Given that an individual tests positive, what is the probability that they are truly infected? Use Bayes' theorem to calculate this quantity which is known as the *positive predictive value* (PPV). Is this a high or low PPV?**

We want to calculate the probability an indivdiual is infected $I$ given that they test postiive $P$, $\Pr(I|P)$. Using Bayes' theorem we have:

$$\Pr(I|P) = \frac{\Pr(P|I)\Pr(I)}{\Pr(P)}$$

From the prevenlence of the infection we have that $\Pr(I) = 0.05$.

From the senstativity of the infection we have that $\Pr(P|I) = 0.9$

But to calculate the probability of a postive test we have to use the law of total probability:

$$\begin{aligned}\Pr(P) =& \Pr(P|I)\Pr(I) + \Pr(P|I')\Pr(I') \\ =& 0.9 \times 0.05 + (1 - 0.85) * (1 - 0.05)\end{aligned}$$

where $I'$ indicated not infected.

```
[1]: (0.9*0.05)/(0.9*0.05+0.15*0.95)
```

```
[1]: 0.24000000000000002
```

The positive predictive value of $\Pr(I|P) = 0.24$ or 24% which is quite low!

Grading: 2pts for equation, 1pt for positive predictive value, 1pt for interpretation

**Part B. [4pts] Given that an individual tests negative, what is the probability that they are truly uninfected? Use Bayes' theorem to calculate this *negative predictive value* (NPV). Is this a high or low NPV?**

We want to calculate $\Pr(I'|P')$. Using Bayes' theorem we have:

$$\Pr(I'|P') = \frac{\overbrace{\Pr(P'|I')}^{0.85}\,\overbrace{\Pr(I')}^{0.95}}{\underbrace{\Pr(P')}_{0.05\times0.1+0.95\times0.85}}$$

```
[2]: 0.85*0.95/(0.85*0.95+0.05*0.1)
```

```
[2]: 0.9938461538461538
```

This is a fairly high negative predictive value.

Grading: 2pts for equation, 1pt for value, 1pt for interpretation

**Part C. [2pts] Discuss the impact of test sensitivity and specificity on the accuracy of diagnostic tests, especially in the context of infectious diseases.**

These results show that for can seem to be a fairly sensitive and specific test that the utility of the test, particularly for finding infected cases, which are inherently rare, is limited. Having 5% of the population infected at any one time is a very high prevalence of the disease, the results will be even more stark for rare infections in which only a fraction of a percent of individuals are infected.

**Part D. [2pts] Suppose that individuals are tested twice. Assuming that the accuracy of the tests is independent, what is the probability that an individual is infected given two negative test results? Discuss possible pros and cons of a multiple-testing design.**

(1pt) We want to know $\Pr(I|P'P')$. Let's use Baye's Theorem.

$$\Pr(I|P'P') = \frac{\Pr(P'P'|I)\,\overbrace{\Pr(I)}^{0.05}}{\Pr(P'P')}$$

Using the independence of tests we have:

$$\Pr(P'P'|I) = \Pr(P'|I)\,\Pr(P'|I) = (1 - \overbrace{\Pr(P|I)}^{0.9})(1 - \Pr(P|I)) = 0.1^2 = 0.01$$

Now the denominator

$$\Pr(P'P') = \Pr(P'P'|I)\Pr(I) + \underbrace{\Pr(P'P'|I')}_{\Pr(P'|I')^2}\Pr(I') = 0.1^2 * 0.05 + 0.85^2 * 0.95$$

This gives:

$$\Pr(I|P'P') = 0.00073$$

```
[1]: 0.1*0.1*0.05/(0.1*0.1*0.05+0.85*0.85*0.95)
```

```
[1]: 0.0007279344858962697
```

(1pt) This dual sampling design dramatically decreases the frequency of false negatives (negative even though you are infected). A con would be cost and the potential reduction in the total number of people tested.

## 1.2 2. Genetic Inheritance and Probability

**In genetics, the principles of probability are often used to understand the inheritance of traits. Consider a specific genetic trait determined by a single gene with two alleles: dominant (D) and recessive (d). In a population, 40% of individuals are homozygous dominant (DD), 30% are heterozygous (Dd), and 30% are homozygous recessive (dd) for this trait.**

**Part A. [3pts] What is the probability that a randomly selected individual exhibits the dominant trait?**

By definition of genetic dominance, the dominant trait will be expressed by both DD and Dd individuals. So the probability is $0.4 + 0.3 = 0.7$ or 70%.

**Part B. [3pts] Given that two individuals heterozygous for this trait (Dd) mate and have offspring, what is the probability that their child will express the dominant trait?**

Drawing the Punnet Square we have:

|   | D | d |
|---|---|---|
| D | DD | Dd |
| d | Dd | dd |

Dominat Trait

75% of the offspring express the dominant trait

**Part C. [3pts] Consider a different co-dominant genetic trait determined by a separate (unlinked) gene. In the focal population, 60% of individuals express the mutant phenotype, 30% express the heterozygous phenotype, and 10% express the ancestral phenotype. What is the probability that a randomly selected individual is heterozygous for the first trait (Dd) and expresses the mutant phenotype (MM) for the second trait?**

By unlinked we mean that the genotype at the first gene and the genotype at the second gene are independent. So we have:

$\Pr(Dd) \times \Pr(MM) = 0.3 \times 0.6 = 0.18$

```
[3]: 0.3*0.6
```

```
[3]: 0.18
```

**Part D. [1pt] Discuss the importance of understanding and applying probabilities in the field of genetics and how it can aid in predicting the likelihood of trait inheritance and genotypic outcomes.**

Probability, as illustrated above, plays an important role in genetic counselling. For example in detecting the probability that an individual will develop a rare recessive disease or in informing population-level differences in the occurrence of a disease such as sickle cell anemia.

### 1.2.1 3. Genetic Inheritance of a Recessive Trait

**In the study of genetics, we often encounter situations where we need to model the probability of specific genetic outcomes. Let's consider a simple gene with two alleles, a recessive (r) allele and a dominant (R) allele. An individual can inherit the allele from either parent. The frequency of the 'r' allele in the population is given by $p = 0.5$.**

**Part A. [3pts] Define a random variable that represents the maternally inherited allele. What are the possible outcomes of this random variable, and what are their probabilities? This random variable is distributed according to what distribution?**

(2pts) The maternally inherited allele can be either an 'r' or an 'R'. Let's define a random variable:

$$X = \begin{cases} 1 & 'r' \\ 0 & 'R' \end{cases}$$

(1pts) We have that $\Pr(X = 1) = 0.25$ and hence $\Pr(X = 0) = 0.75$. This is a Bernoulli random variable and is distributed like $X \sim \mathcal{Ber}(0.25)$.

**Part B. [4pts] Explain how you can model the inheritance of a recessive allele (r) from both parents using a Binomial distribution. What does the r.v. represent in this case? What are the parameters of the Binomial distribution in this context, and what do they represent?**

(2pts) Let's define a r.v. giving the number of 'r' alleles inherited from both parents.

$$X = \begin{cases} 0 & 'RR' \\ 1 & 'Rr' \text{ or } 'rR' \\ 2 & 'rr' \end{cases}$$

(2pts) This r.v. is distributed according to $X \sim \mathcal{B}er(N = 2, p = 0.25)$ where $N$ is the number of parents/chromosomes and $p$ is the 'success' probability of inheriting the recessive allele.

**Part C. [3pts] Calculate the probability that an individual inherits two recessive alleles (rr), one from each from both parents for this trait.**

We have:

$$\Pr(X = 2) = \binom{2}{2} 0.25^2 (1 - 0.25)^{(2-2)} = 0.125$$

**4. Moment Analysis in Measuring Tree Height in a Forest In forest ecology, understanding the height distribution of trees is crucial for assessing forest structure and biomass. Ecologists often use probability distributions to describe the distribution of tree heights. Let's consider a forest where tree heights follow a continuous probability distribution, f(h), where h is the height of a tree in meters.**

**The probability density function for tree heights in this forest is given by:**

$$f(h) = khe^{-0.2h} \quad \text{for } 0 \le h < \infty$$

**Part A: [1pts] What is the value of where $k$?**

Since $f(h)$ is a probability distribution it must sum to 1 hence we have:

$$\int_0^\infty khe^{-0.2h} dh = 1$$

Using integration by parts ($u = h \quad v = \frac{1}{-0.2} e^{-0.2h}$) we have:

$$k \left( he^{-0.2h} \big|_0^\infty - \int_0^\infty \frac{1}{-0.2} e^{-0.2h} dh \right)$$

$$k \left( 0 - \frac{1}{-0.04} \right) = k * 25$$

So $k = \frac{1}{25}$

**Part B: [2pts] Calculate the mean tree height.**

Mean height is given by:

$$E[h] = \int_0^\infty \frac{h^2}{25} e^{-0.2h} dh = \frac{1}{25} e^{-0.2h} \left( -5.h^2 - 50.h - 250. \right) \Big|_0^\infty = 10$$

**Part C: [2pts] Calculate the variance in tree height distribution.**

The variance equals $Var[h] = E[h^2] - E[h]^2$

Finding $E[h^2] = \int_0^\infty \frac{h^3}{25} e^{-0.2h} dh = \frac{1}{25} e^{-\frac{h}{5}} \left(-5h^3 - 75h^2 - 750h - 3750\right)\Big|_0^\infty = 150$

Hence the variance is $150 - 10^2 = 50$

**Part D: [3pts] Calculate the third moment (skewness) of the tree height distribution. Discuss what the answers to A,B, and C tell you about the tree height distribution.**

(2pt) $Skew[h] = E[h^3] - 3E[h^2]\mu + 2\mu^3$ where $\mu = E[h]$

We have:

$ E[hˆ3]=3000 $

[30]:
```
3000-3*150*10+2*10**3
```

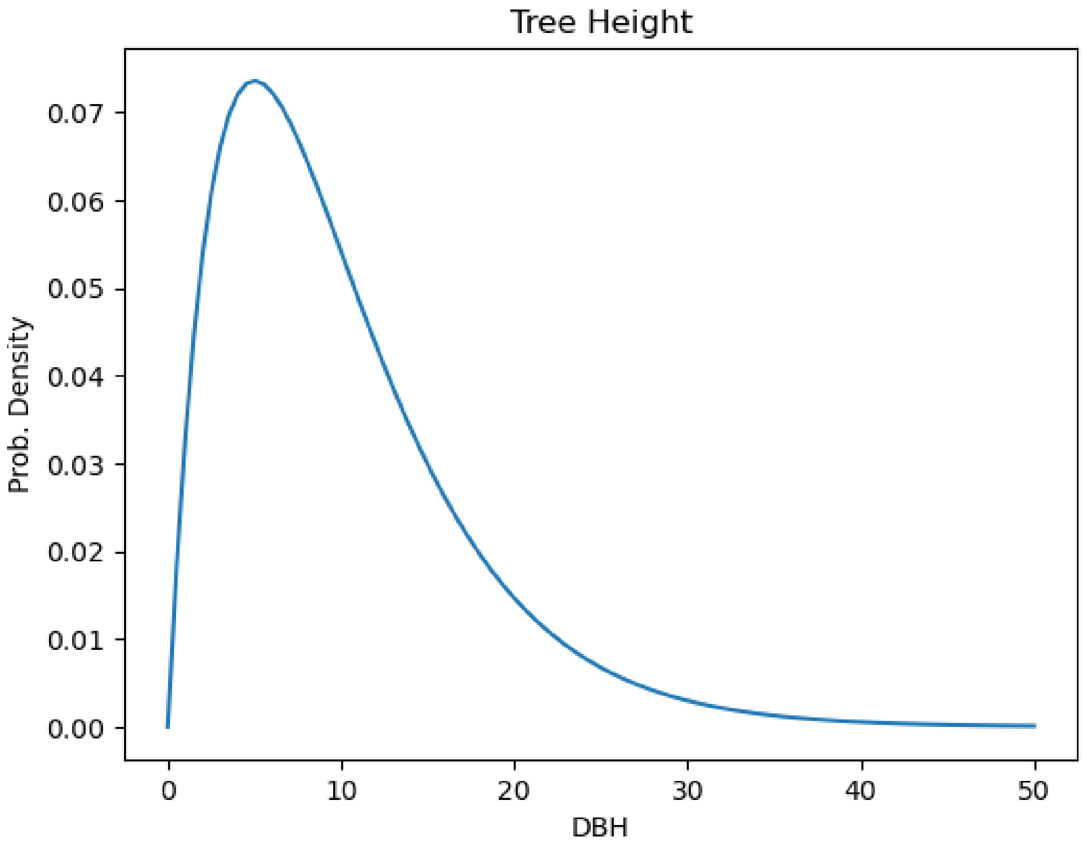
[30]:
```
500
```

$$Skew[h] = 500$$

(1pt) The distribution is **right skewed**. This means that the mean is less than the mode. So most trees have a DBH$<$ 10. Given the mean and variance, there is a long right tail (large trees). So most trees are small but a few are very large.

**Part E: (2pt) Plot the distribution of tree heights**

[36]:
```
x_values = np.linspace(0, 50, 100)
y_values=x_values*np.exp(x_values*(-0.2))/25
plt.plot(x_values, y_values)
plt.xlabel('DBH')
plt.ylabel('Prob. Density')
plt.title('Tree Height');
```



**Part F: (1pt) Suppose you want to compare the tree height distribution of this forest to another forest. How could you use moments (mean, variance, skewness) to assess and compare the ecological characteristics of these two forests based on tree height data?**

A comparison of the means will tell us about the total biomass in the forest. A comparison of the variances will tell us about the amount of diversity in ages/health of the trees and the skew will tell us about how many very large trees there are for example for logging.

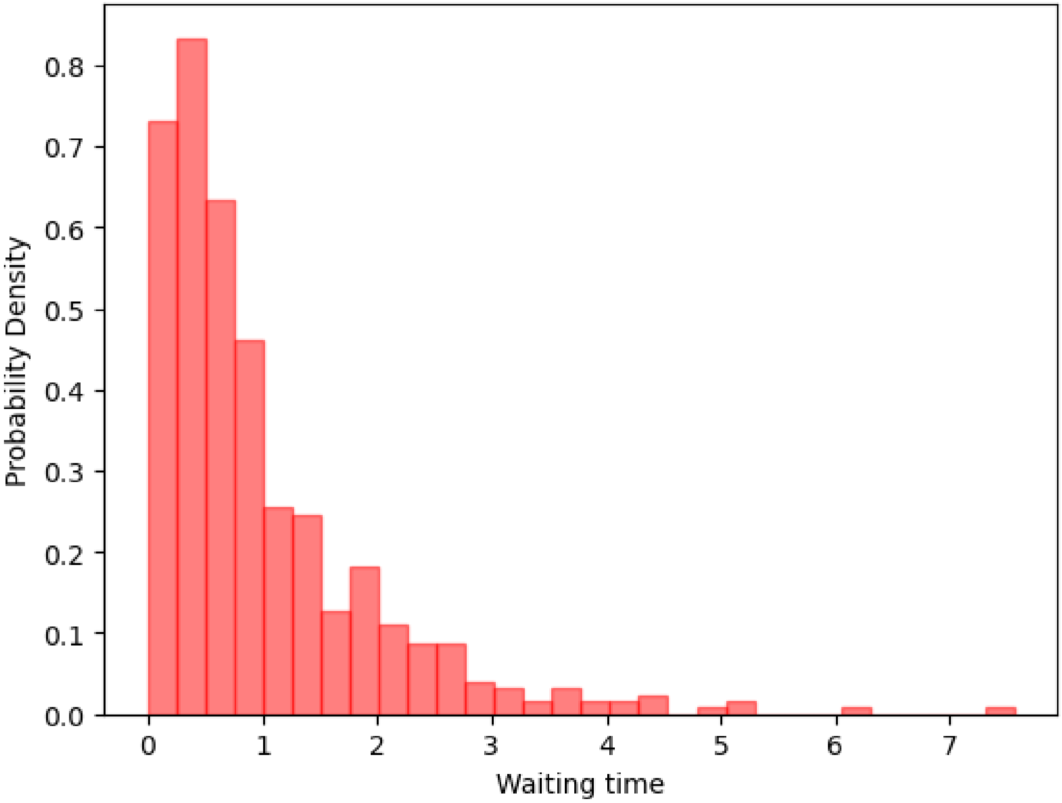
**5. Random number generation Part A: (4pt) Modify the python code from class to generate a histogram of 500 exponentially distributed random numbers where $\lambda = 1$. Show the resulting plot.**

```
[3]: lam=1
     from scipy.stats import expon
     uList = np.random.rand(500)
     xList=expon.ppf(uList, scale=1/lam); #distriution is specified in terms of␣
       ↪scale=1/lambda
     np.mean(xList)
```

```
[3]: 0.9792342456503971
```

[4]:

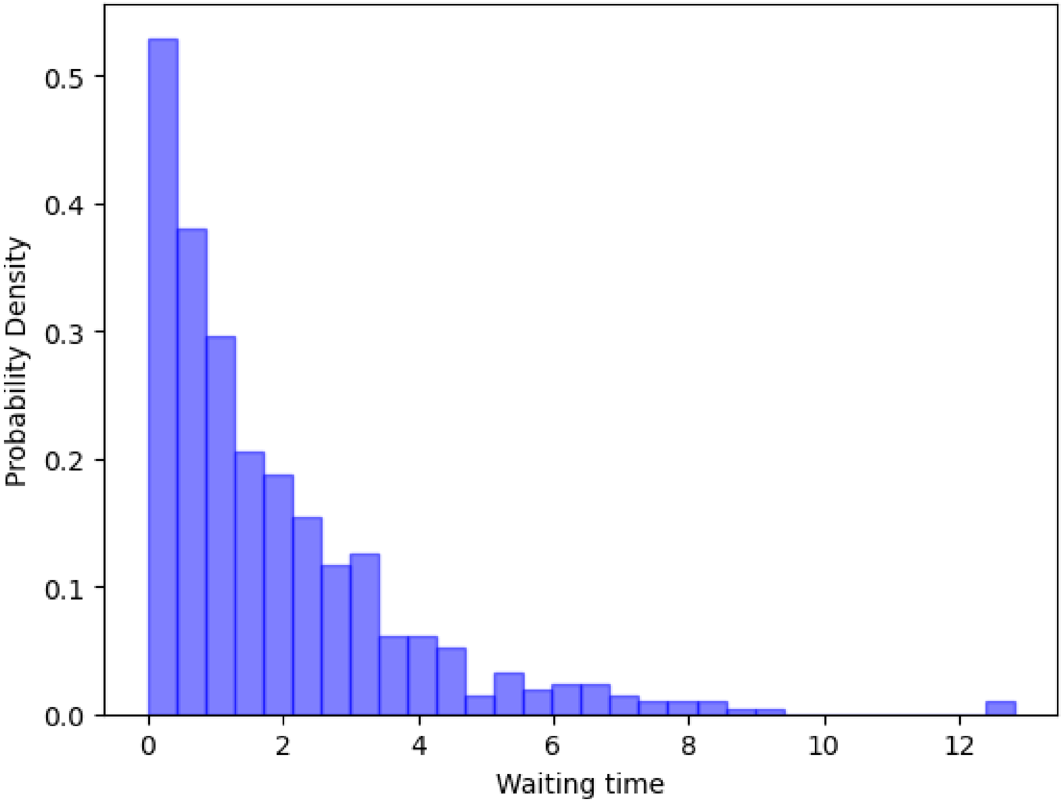
```
plt.hist(xList, bins=30, density=True, alpha=0.5, color='red', edgecolor='red');
plt.xlabel('Waiting time')
plt.ylabel('Probability Density');
```



**Part B: (4pt) Use the built in function *np.random.exponential* to generate 500 random numbers from an exponential distribution where $\lambda = \frac{1}{2}$. Show the resulting plot.**

[5]:

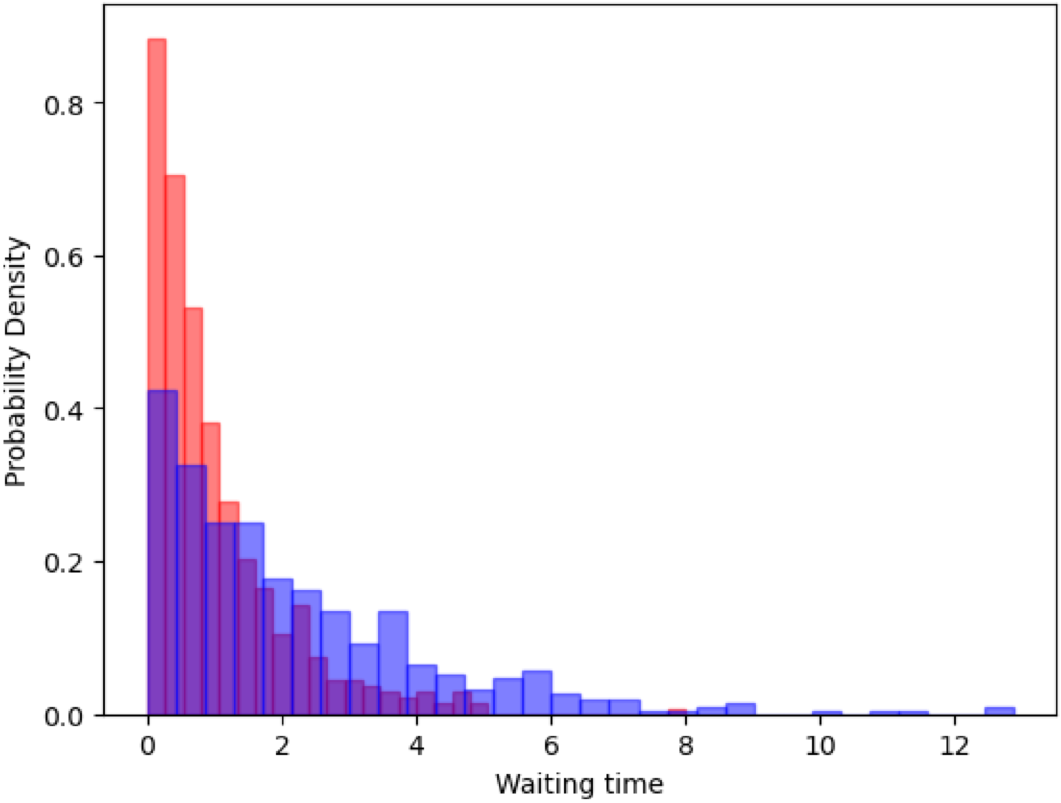
```
lam2=0.5
random_numbers = np.random.exponential(scale=1/lam2, size=500)
plt.hist(random_numbers, bins=30, density=True, alpha=0.5, color='blue',␣
  ↪edgecolor='blue');
plt.xlabel('Waiting time')
plt.ylabel('Probability Density');
```



**Part C: (2pt) Plot the distributions from part A and part B on the same plot and use the result to compare the two distributions.**

Grading: 1pt for plot, 1pt for comparison

[48]:
```
plt.hist(xList, bins=30, density=True, alpha=0.5, color='red', edgecolor='red');
plt.hist(random_numbers, bins=30, density=True, alpha=0.5, color='blue',␣
  ↪edgecolor='blue');
plt.xlabel('Waiting time')
plt.ylabel('Probability Density');
```



Blue distribution has a larger mean and hence a lower probability of short waiting times and a longer tail.

**6. Challenge Question: Modelling Disease Incubation Periods Understanding the incubation period of a disease is crucial in epidemiology, as it helps assess disease transmission dynamics and develop effective control measures. Let's consider a new infectious disease with a known incubation period, $T$ during which individuals are infected, asymptomatic, but still infectious. The incubation period, denoted as follows a continuous probability distribution with the following probability density function (PDF):**

$$\Pr(T = t) = f(t) = 0.2e^{-0.2t} \quad \text{for } t \geq 0$$

**Part A. [3pt] Calculate the cumulative density function (CDF) of the incubation period, $F(t)$. What does the CDF represent in the context of this infectious disease, and how can it be useful for epidemiologists?**

(2pt) By inspection we have that this is an exponential distribution with rate $\lambda = 0.2$ so the CDF is known to be:

$$F(t) = 1 - e^{-0.2t}$$

(1pt) This distribution tells us the probability that an individual is already symptomatic after being infected for $t$ units of time.

**Part B. [3pt] Calculate the probability that an individual exposed to the disease will develop symptoms within the first 5 days (0 t 5). Use the CDF to find this probability.**

This probabilitly is given by $F(5)$

```
[49]: 1-math.exp(-0.2*5)
```

```
[49]: 0.6321205588285577
```

**Part C. [3pt] Calculate the mean and skew of the incubation period using the PDF. What do these movements mean to imply about disease transmission and control strategies??**

(1pt) The mean of an exponential distribution is simply $\frac{1}{\lambda} = 1/0.2 = 5$
(1pt) The skew of the exponential distribution is: 2

(1pt) This is how long we would "expect" it to take for a random individual to develop symptoms. The large mean implies that there will be 5 days during which individuals can have 'hidden transmission' where they are transmitting the infection but aren't yet feeling ill. Note however that the distribution is right-skewed (the mean is larger than the mode). So while it is our best guess for the time until any one individual is 5 days infected most individuals will develop symptoms sooner.

**Part D. [1pt] Discuss the epidemiological relevance of using probability density and cumulative density functions to model disease incubation periods. How can understanding these functions help in predicting disease spread, estimating outbreak sizes, and planning public health interventions?**

The PDF and CDF of the incubation period help us understand the efficacy of tests, for example for contact tracing, when those tests require people to be symptomatic.