Assignment 2 - Solutions

Question 1. A Discrete Distribution - PMF

Verify that $p(x)$ is a probability mass function (*pmf*) and calculate the following for a random variable X with this pmf:

x	1.25	1.5	1.75	2	2.45
$p(x)$	0.25	0.35	0.1	0.15	0.15

To verify that $p(x)$ is a probability mass function sum p(x) terms to see they equal 1.

$$\sum_x p(x) = 0.25 + 0.35 + 0.1 + 0.15 + 0.15 = 1$$

To confirm this in Julia we need to define two vectors and then sum the probabilities.

In [1]:
```
x = [1.25,1.5,1.75,2,2.45];
px = [0.25,0.35,0.1,0.15,0.15];
sum(px)
```

Out[1]: 1.0

(a) $P(X \leq 2)$

Here sum the entries of $p(x)$ that relate to events less than or equal to 2.

$$P(X \leq 2) = \sum_{x=1.25}^{2} p(x) = 0.25 + 0.35 + 0.1 + 0.15 = 0.85$$

(b) $P(X > 1.65)$

Again sum the entries of $p(x)$ that relate to events greater than 1.65, that is 1.75, 2 and 2.45

$$P(X > 1.65) = \sum_{x=1.75}^{2.45} p(x) = 0.1 + 0.15 + 0.15 = 0.4$$

(c) $P(X = 1.5)$

As this is a discrete distribution, just read off from the pmf $P(X = 1.5) = 0.35$.

(d) $P(X < 1.3 \text{ or } X \geq 21)$

As 21 is past the end of the pmf so is not possible, this can be rewritten as $P(X < 1.3)$ and solved as with the questions above

$$P(X < 1.3) = P(X \leq 1.25) = 0.25$$

(e) The mean.

The mean is equal to the expected value which is simply the sum of the values of x multiplied by their respective probabilities

$$\mu = E(X) = \sum_x xp(x) = 1.25 \times 0.25 + 1.5 \times 0.35 + 1.75 \times 0.1 + 2 \times 0.15 + 2.45 \times 0.15$$

Now calculating this using Julia as a calculator:

In [2]:
```
mean1= sum(x.*px)
```

Out[2]:
```
1.68
```

(f) The variance.

To make the calculation easier first look at the equation for calculating the variance and then rearrange

$$\begin{aligned}\sigma^2 = Var(X) &= \sum_x (x-\mu)^2 p(x) \\ &= \sum_x (x - E(X))^2 p(x) \\ &= \sum_x (x^2 - 2xE(X) + E(X)^2)p(x) \\ &= \sum_x x^2 p(x) - 2E(X)\sum_x xp(x) + E(X)^2 \sum_x p(x) \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= E(X^2) - E(X)^2\end{aligned}$$

As before using Julia as a calculator

In [3]:
```
var1 = sum(x.^2.*px)-mean1^2
```

Out[3]:
```
0.16235000000000044
```

(g) Sketch the cumulative distribution function (*cdf*). Note that it exhibits jumps and is a right continuous function.

To sketch the cumulative distribution function we use the following code

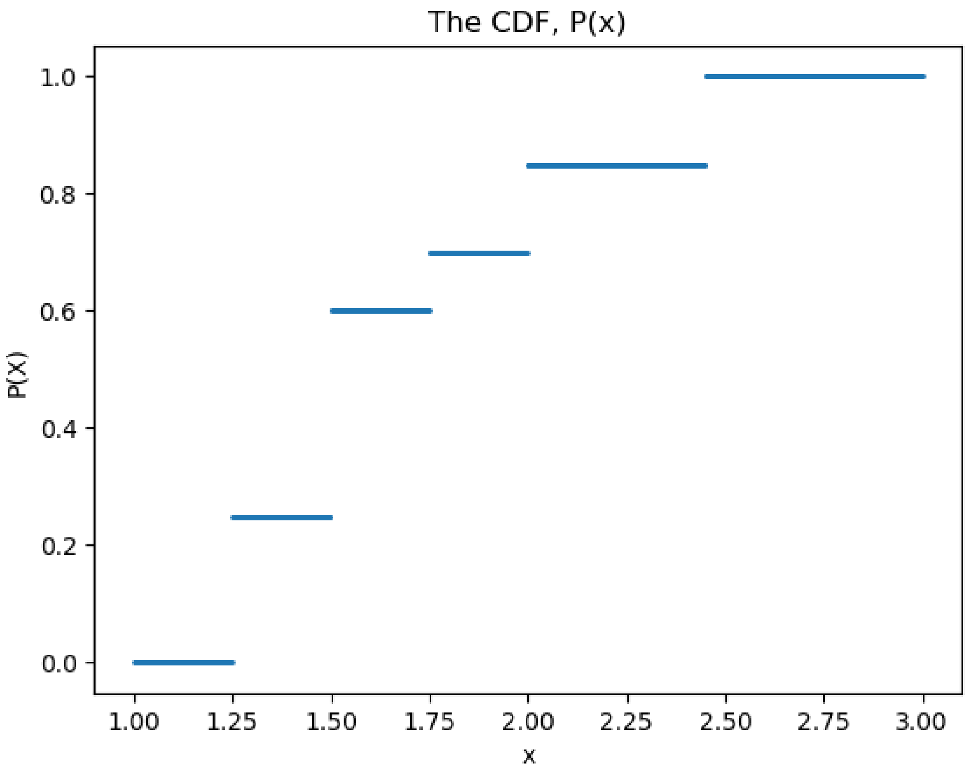
In [4]:
```
using PyPlot

function Question1(x)
    if x < 1.25
        return 0;
    elseif x < 1.5
        return 0.25
    elseif x < 1.75
        return 0.25+0.35
    elseif x < 2
        return 0.25+0.35+0.1
    elseif x < 2.45
        return 0.25+0.35+0.1+0.15
    else
        return 1
    end
end

domain= linspace(1,3,1001)

range= map(x->Question1(x),domain)

PyPlot.scatter(domain,range,s=1);
xlabel("x");
ylabel("P(X)");
title("The CDF, P(x)");
```



Question 2. A Discrete Distribution - CDF

Given the cdf, $F(x)$ below for the random variable, X, calculate the following:

$$F(x) = \begin{cases} 0 & x < -10 \\ 0.45 & -10 \leq x < 45 \\ 0.75 & 45 \leq x < 60 \\ 1 & 60 \leq x \end{cases}$$

(a) $P(X > 50)$

The cdf gives the probability of values less than a value so first find $P(X < 50)$ then subtract from 1. Looking at the cdf 50 falls into the $45 \leq x < 60$ condition so $P(X < 50) = 0.75$. Therefore:

$$P(X > 50) = 1 - P(X < 50) = 1 - 0.75 = 0.25$$

(b) $P(X \leq 40)$

This time the probability can be obtained directly from the cdf. 40 falls into the $-10 \leq x < 45$ condition so

$$P(X \leq 40) = 0.45$$

(c) $P(40 \leq X \leq 90)$

Here a range is being used, so the difference between the start point and the end point needs to be found. 90 is above the top condition so will have a cdf value of 1, 40 as seen above will be 0.45. So the probability of being in the range $40 \leq q \leq 90$ will be:

$$P(40 \leq X \leq 90) = P(X \leq 90) - P(X \leq 40) = 1 - 0.45 = 0.55$$

(d) $P(X < 0)$

Here the condition that is being examined is strictly less than. However as 0 is not on the boundary of any of the conditions in the cdf, the value can be taken from where it lies

$$P(X < 0) = 0.45$$

(e) $P(0 \leq X < 10)$

Again this is a range of values so identify where the start and end points are. As both 0 and 10 are in the same condition the probability of being in that range is 0.

$$P(0 \leq X < 10) = P(X < 10) - P(X \leq 0) = 0.45 - 0.45 = 0$$

(f) $P(-10 < X < 10)$

Here again is a range of values. As the condition is looking for strictly greater than -10 and strictly less than 10 both the start and end points are in the same condition of the cdf, therefore,

$$P(-10 < X < 10) = P(X < 10) - P(X < -10) = 0.45 - 0.45 = 0$$

(g) The mean.

To calculate the mean, first compute the pdf. This is done by identifying the points where the cdf changes, -10, 45 and 60. Taking the probability slightly before and slightly after calculate the probabilities as above. This leads to the probability mass function of

x	-10	45	60
$f(x)$	0.45	0.30	0.25

To aid calculations input these vectors into Julia. As seen above the mean can be computed by taking the sum of each value multiplied by its probability

```
In [5]: x2 = [-10,45,60]
        fx = [0.45,0.30,0.25]
        mean2=sum(x2.*fx)
```

```
Out[5]: 24.0
```

(h) The standard deviation.

Again following the example in Question 1 the variance can be computed. Remember that the standard deviation is the square root of the variance (so it has the same units as the mean) so square root the computed variance. This is computed in Julia as:

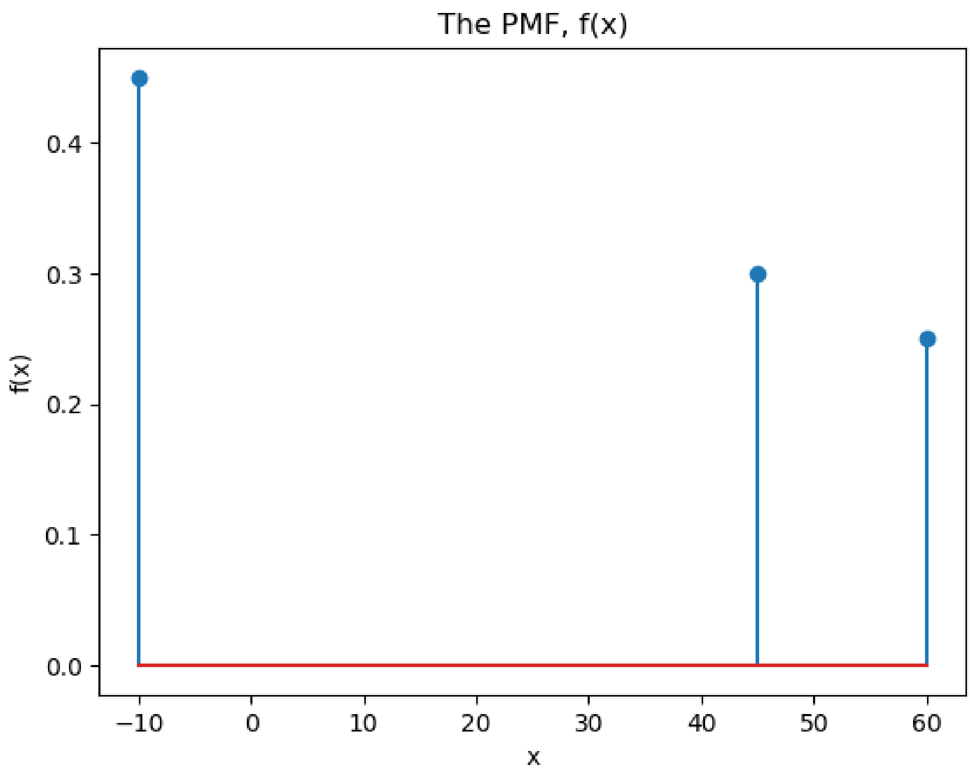
In [6]:
```
sqrt(sum(x2.^2.*fx)-mean2^2)
```

Out[6]:
```
31.24899998399949
```

(i) Sketch the probability mass function (PMF).13:31-15:25 15:35

To sketch the probability mass function just draw a stem plot of x versus f(x) as shown below

In [7]:
```
PyPlot.stem(x2,fx);
xlabel("x");
ylabel("f(x)");
title("The PMF, f(x)");
```



Question 3. Guessing on Multiple Choice Exams

A multiple-choice test contains 25 questions, each with 3 answers. Assume that a student just guesses on each question.

(a) What is the probability that the student answers more than 15 questions correctly?

The number of correct answers is distributed with a Binomial Distribution with $n = 25$ and $p = \frac{1}{3}$. As this is a discrete distribution $P(X > 15) = P(X \geq 16)$ so the probabilities of getting 16 to 25 answers correct need to be totalled.

$$\begin{aligned}P(X \geq 16) &= P(X = 16) + P(X = 17) + \ldots + P(X = 25)\\ &= \binom{25}{16}\left(\frac{1}{3}\right)^{16}\left(\frac{2}{3}\right)^{9} + \binom{25}{17}\left(\frac{1}{3}\right)^{17}\left(\frac{2}{3}\right)^{8} + \ldots + \binom{25}{25}\left(\frac{1}{3}\right)^{25}\left(\frac{2}{3}\right)^{0}\\ &= 0.00123453 + 0.00032679 + \ldots + 1.180235 \times 10^{-12}\\ &= 0.01649583\end{aligned}$$

(b) What is the probability that the student answers fewer than 10 questions correctly?

Here again as the questions is asking for strictly less than 10 questions correct and following a similar method to that of the previous question

$$\begin{aligned}
P(X \leq 9) &= P(X=9) + P(X=8) + \ldots + P(X=0) \\
&= \binom{25}{9}\left(\frac{1}{3}\right)^9\left(\frac{2}{3}\right)^{16} + \binom{25}{8}\left(\frac{1}{3}\right)^8\left(\frac{2}{3}\right)^{18} + \ldots + \binom{25}{25}\left(\frac{1}{3}\right)^0\left(\frac{2}{3}\right)^{25} \\
&= 0.1580198 + 0.1673151 + \ldots + 3.960213 \times 10^{-05} \\
&= 0.6955988
\end{aligned}$$

The following code generates the vector `pmfValues` from a Binomial distribution with parameters $n = 10$ and $p = 0.4$. It then sums up the vector, illustrating that the sum of all of the probabilities is 1.

In [8]:
```
using Distributions
bDist = Binomial(10 ,0.4)
pmfValues = [pdf(bDist , x) for x in 0:10]
sum(pmfValues[1:11])
```

Out[8]: 1.0

(c) Modify the code above, to validate your answers in (a) and (b).

In [9]:
```
#part (a)
bDist = Binomial(25, 1/3)
pmfValues = [pdf(bDist , x) for x in 0:25]
sum(pmfValues[17:26]) # 16 to 25
```

Out[9]: 0.0016495826751628548

In [10]:
```
#part (b)
sum(pmfValues[1:10]) # 0 to 9
```

Out[10]: 0.6955987713554281

Question 4. Stuck in Traffic

A particularly long traffic light on your morning commute is green 35% of the time that you approach it. Assume that each morning represents an independent trial.

(a) Over 3 mornings, what is the probability that the light is green on exactly one day?

The green lights can be considered as a Binomial distributed random variable with $p = 0.35$. For this part of the question $n = 3$, so calculate $P(G = 1)$ by

$$\begin{aligned}
P(G=1) &= \binom{3}{1} 0.35^1 0.65^2 \\
&= \frac{3!}{2!1!} 0.35 \times 0.65^2 \\
&= 0.443625
\end{aligned}$$

(b) Over 15 mornings, what is the probability that the light is green on exactly four days?

Now $n = 15$ with $p = 0.35$ still and calculate $P(G = 4)$ by

$$\begin{aligned} P(G = 4) &= \binom{15}{4} 0.35^4 0.65^{11} \\ &= \frac{15!}{4!11!} 0.35^4 \times 0.65^{11} \\ &= 0.1792469 \end{aligned}$$

(c) Over 15 mornings, what is the probability that the light is green on more than four days?

Here $P(G > 4)$ is being asked for, so calculate $P(G \geq 5)$.

$$\begin{aligned} P(G > 4) = P(G \geq 5) &= 1 - P(G \leq 4) \\ &= 1 - P(G = 4) - P(G = 3) - P(G = 2) - P(G = 1) - P(G = 0) \\ &= 1 - 0.1792469 - 0.1109624 - 0.04755531 - 0.01261672 - 0.001562069 \\ &= 0.6480566 \end{aligned}$$

(d) What is the mean number of days with green light during a month of 30 days?

To calculate the mean of a Binomial random variable, multiply the number of trials by the probability of the event. Hence

$$\mu = E(G) = 30 \times 0.35 = 10.5$$

Optional: You may verify your analytic answers using Julia, in a similar manner to Question 3c.

In [11]:
```
#part (a)
bDist4a = Binomial(3,0.35)
pdf(bDist4a,1)
```

Out[11]: 0.443625

In [12]:
```
#part (b)
bDist4bc = Binomial(15,0.35)
pdf(bDist4bc,4)
```

Out[12]: 0.17924694060678328

In [13]:
```
#part (c)
ccdf(bDist4bc,4) # compliment of cumulative distribution function
```

Out[13]: 0.6480565722237326

In [14]:
```
#part (d)
30*0.35
```

Out[14]: 10.5

Question 5. Aerospace Inspections

The thickness of a flange on an aircraft component is Uniformly distributed between 0.9 and 1.1 millimetres. Determine the following:

(a) Cumulative distribution function of flange thickness.

First the probability density function of the flanges needs to be determined. This is done by dividing one by the range of the values

$$f(x) = \frac{1}{1.1 - 0.9} = \frac{1}{0.2} = 5$$

Now integrate the pdf to get the cumulative distribution function

$$F(x) = \int_{-\infty}^{x} 5\ du = \int_{0.9}^{x} 5\ du = 5(x - 0.9)$$

(b) Proportion of flanges that exceeds 1.02 millimetres.

The cdf above gives $P(X < x)$ so to calculate the proportion of flanges that exceeds 1.02 millimetres the following calculation is performed

$$\begin{aligned} P(X > 1.02) &= 1 - P(X < 1.02) \\ &= 1 - F(1.02) \\ &= 1 - 5(1.02 - 0.9) \\ &= 0.4 \end{aligned}$$

(c) Thickness exceeds 75% of the flanges.

This is the 0.75 quantile so the following equation is obtained

$$0.75 = 5(x - 0.9)$$

Solving this equation for x

$$\begin{aligned} 0.75 &= 5(x - 0.9) \\ 0.15 &= x - 0.9 \\ 0.15 + 0.9 &= x \\ 1.05 &= \end{aligned}$$

So the thickness needs to be at least 1.05 millimetres.

(d) Mean and standard deviation of flange thickness.

To calculate the mean the equation given in the course lecture notes or $\int_{-\infty}^{\infty} x f(x)\ dx$ can be used. However an alternative method is to set $U \sim \text{Uniform}(0, 1)$. Then $X = 0.2U + 0.9$. For U it is know that $E(X) = \frac{1}{2}$ and $Var(U) = \frac{1}{12}$, so $E(X) = 0.2E(U) + 0.9 = 1$. Further $Var(X) = 0.2^2 Var(U) = \frac{0.04}{12} = 0.00\dot{3}$.

So the mean is $\mu = 1$ mm and the standard deviation, which is the square root of the variance, is $\sigma = \sqrt{0.00\dot{3}} = 0.05774$ mm.

(e) Assume now that you are sampling 12 independent flanges. What is the variance of the number of flanges with a thickness less than 0.96 millimetres?

First calculate the probability of obtaining a flange with a thickness less than 0.96 millimetres

$$P(X < 0.96) = F(0.96) = 5(0.96 - 0.9) = 0.3.$$

The number of flanges from the 12 independent samples can be considered to be a Binomial distributed variable with $n = 12$ and $p = 0.3$. The variance of a Binomial variable is $\sigma^2 = np(1 - p)$ therefore $\sigma^2 = 12 \times 0.3 \times 0.7 = 2.52.$

Question 6. Mobile Phone Semiconductors

The line width for semiconductor manufacturing is assumed to be Normally distributed with a mean of 0.7 micrometers and a standard deviation of 0.06 micrometers.

(a) What is the probability that a line width is greater than 0.72 micrometer?

Here the standard normal tables from the exam booklet will be used. These tables give the lower tail probability (i.e. $P(Z < z)$) of the z-values. So the compliment of the event $P(X < 0.72)$ will equal $P(X > 0.72)$. Also this value needs to be standardised so that it can be found in the table. The calculation is then:

$$\begin{aligned} P(X > 0.72) &= 1 - P(X < 0.72) \\ &= 1 - P\left(Z < \frac{0.72 - 0.7}{0.06}\right) \\ &= 1 - P\left(Z < 0.\dot{3}\right) \\ &= 1 - 0.6306 = 0.3694 \end{aligned}$$

(b) What is the probability that a line width is between 0.57 and 0.67 micrometer?

To calculate this probability find the difference between $P(X < 0.67)$ and $P(X < 0.57)$

$$\begin{aligned} P(0.57 < X < 0.67) &= P(X < 0.67) - P(X < 0.57) \\ &= P\left(Z < \frac{0.67 - 0.7}{0.06}\right) - P\left(Z < \frac{0.57 - 0.7}{0.06}\right) \\ &= 0.3085 - 0.0151 \\ &= 0.2934 \end{aligned}$$

(c) The line width of 80% of samples is below what value?

This is the 80% quantile and $P(Z < z) = 0.8$ when $z = 0.8416$. Enter the know values for μ, σ and z into the standardisation formula and solve for x.

$$\begin{aligned} 0.8416 &= \frac{x - 0.7}{0.06} \\ 0.0504 &= x - 0.7 \\ 0.7505 &= x \end{aligned}$$

Therefore the line width would need to be 0.7505 micrometres.

Question 7. The Prototype Shoe

The weight of a sophisticated running shoe is normally distributed with a mean of 10 ounces and a standard deviation of 0.7 ounce.

(a) What is the probability that a shoe weighs more than 13 ounces?

This probability is the compliment of $P(X < 13)$ and so is calculated as follows.

$$\begin{aligned} P(X > 13) &= 1 - P(X < 13) \\ &= 1 - P\left(Z < \frac{13 - 10}{0.7}\right) \\ &= 1 - P(Z < 4.29) \\ &< 1 - 0.9999 \\ &< 0.00005 \end{aligned}$$

As $P(Z < 4.29) > 0.9999.$

(b) What must the standard deviation of weights be in order for the company to state that 99% of its shoes weighs less than 13 ounces?

This is the 99% quantile so $P(Z < z) = 0.99$ when $z = 2.326348$. Now take the standardisation formula and enter known values for x μ and z. Rearrange to solve for σ

$$\begin{aligned} 2.33 &= \frac{13 - 10}{\sigma} \\ \sigma &= \frac{3}{2.33} \\ &= 1.28 \quad \text{actual } 1.289575 \end{aligned}$$

(c) If the standard deviation remains at 0.7 ounce, what must the mean weight be for the company to state that 99% of its shoes weighs less than 13 ounces?

Again as $z = 2.326348$ for the 99% quantile, substitute in known values for σ, z and x and rearrange for μ

$$\begin{aligned} 2.33 &= \frac{13 - \mu}{0.7} \\ 1.628 &= 13 - \mu \\ \mu &= 11.369 \quad \text{actual } 11.37156 \end{aligned}$$

Question 8. Time Until (Blue Screen of Death) BSoD

Suppose that the time to failure (in hours) of hard drives in a personal computer can be modelled by an exponential distribution with $\lambda = 0.002$.

(a) What proportion of the hard drives will last at least 8,000 hours?

For an exponentially distributed variable $P(X > x) = 1 - F(x) = e^{-\lambda x}$.Using this the calculation is as follows

$$\begin{aligned} P(X \geq 8000) &= 1 - F(8000) \\ &= e^{-0.002 \times 8000} \\ &= 1.125352 \times 10^{-07} \end{aligned}$$

So the proportion of hard drives that will last at least 8000 hours is 1.124×10^{-7}.

(b) What proportion of the hard drives will last at most 7,000 hours? Again using the above equation and remembering that $P(X < 7000) = 1 - P(X > 7000)$ we get the following:

$$P(X \leq 7000) = 1 - e^{-0.002 \times 7000} = 0.9999992$$

So the proportion of hard drives that will last at most 7000 hours is > 0.999

(c) What is the variance of the time until failure for a hard drive?

$$Var(X) = \frac{1}{0.002^2} = 250000$$

(d) Use Monte Carlo simulation to predict the following: Assume a computer now has three independent hard-drives and the failure of the computer occurs once all three hard-drives have died. What is the mean life of the computer?

To perform this simulation in Julia, first define a random variable with a Exponential distribution with $\lambda = 0.002$ remembering that Julia will need `Exponential(1/0.002)`. Then create a matrix of random numbers distributed with the distribution using `rand(distribution,iterations,no.drives)`. Use the `maximum` function to perform a row by row max on this matrix and then take the mean of the resulting vector.

In [15]:
```
dist8 = Exponential(1/0.002)
mean(maximum(rand(dist8,10^8,3),2))
```

Out[15]: 916.6399363082636

Alternatively using a `for` loop:

In [16]:
```
maxlife= [max(rand(dist8),rand(dist8),rand(dist8)) for _ in 1:10^8]
mean(maxlife)
```

Out[16]: 916.5980557828449

Without using the distribution package it can be noted that $-\frac{1}{\lambda}\log(U)$ is distributed like $\exp\lambda$ when U is Uniform(0,1)

In [17]:
```
meanLifeEstimate = mean(maximum(-1./0.002.*log.(rand(10^8,3)),2))
```

Out[17]: 916.7097718242395