1. ISLR chapter 5, exercise 2 (page 197-198)

1. We will now derive the probability that a given observation is part of a bootstrap sample. Suppose that we obtain a bootstrap sample from a set of n observations.

(a) What is the probability that the first bootstrap observation is

not the jth observation from the original sample? Justify your answer.

1−1/n.

(b) What is the probability that the second bootstrap observation

is not the jth observation from the original sample?

Since we draw with replacement, it is the same as above. So, 1−1/n.

(c) Argue that the probability that the jth observation is not in the bootstrap sample is (1 – 1/n)n.

With replacement, the probability of the jth observation is not in the bootstrap sample but the product of the probabilities, so (1−1/n)⋯(1−1/n)=(1−1/n)n, where as these probabilities are independant.

(d) When n = 5, what is the probability that the jth observation is in the bootstrap sample?

By pluging into the formular, P(5th observation)=1−(1−1/5)^5=0.672.

(e) When n = 100, what is the probability that the jth observation

P(100th observation)=1−(1−1/100)^100=0.634.

(f) When n = 10, 000, what is the probability that the jth observation

is in the bootstrap sample?

P(10, 000th observation)=1−(1−1/10000)^10000=0.632.

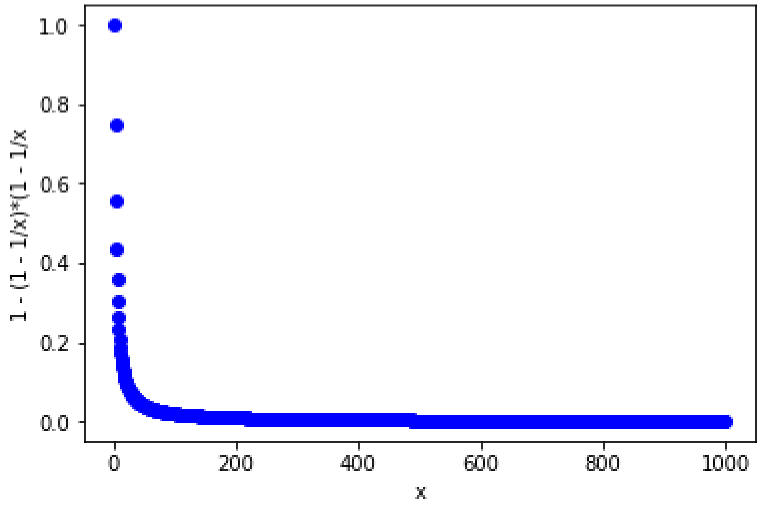
(g) Create a plot that displays, for each integer value of n from 1 to 100, 000, the probability that the jth observation is in the bootstrap sample. Comment on what you observe.

I observe that the plot quickly reaches an asymptote at about 0.632.

In [2]:

```
import matplotlib.pyplot as plt
for x in range (1,100000):
    x = float(x)
    y = float(1 - (1 - 1/x)*(1 - 1/x))
    plt.plot(x, y, 'bo')
    plt.xlabel('x')
    plt.ylabel('1 - (1 - 1/x)*(1 - 1/x')
```

executed in 7.15s, finished 15:49:55 2018-11-06



(h) We will now investigate numerically the probability that a bootstrap sample of size n=100 contains the jth observation. Here j=4. We repeatedly create bootstrap samples, and each time we record whether or not the fourth observation is contained in the bootstrap sample.

In [63]:

```
import pandas as pd
import numpy as np
store = []
for x in range (10000):
    s = pd.Series(np.random.randn(100))
    temp = s.sample(n=4,replace=True)
    store.append(1.0) if sum(temp)>0 else store.append(0.0)

print"the probability is ", np.mean(store)

```

executed in 3.29s, finished 16:36:20 2018-11-06

```
the probability is  0.635
```

2 coding part

In [74]:

```
import matplotlib.pyplot as plt
import pandas as pd
import numpy as np
#read data
df = pd.read_csv("/Users/xuzhaokai/Desktop/109 HW4/hw3_divseq_data.csv")
data = df.values # pd dataFrate to matrix
print data.shape
print data

Lars2 = data[:,0]
Malat1 = data[:,1]
mature = data[:,2]
```

executed in 32ms, finished 16:43:58 2018-11-06

```
(817, 3)
[[ 9.95  6.69  1.  ]
 [10.54  8.53  1.  ]
 [ 6.58  8.74  1.  ]
 ...
 [ 3.98  6.51  0.  ]
 [ 4.9   6.16  0.  ]
 [ 3.38  4.95  0.  ]]
```

2. coding part