In [1]:
```
# DSC530-T302
# Stephen Smitshoek
# Week05
# Exercise 6-1
```

In [2]:
```
import numpy as np
import warnings
import hinc
import math
import scipy.stats
import thinkplot
import thinkstats2

warnings.filterwarnings('ignore')
```

In [3]:
```
def InterpolateSample(df, log_upper=6.0):
    """Makes a sample of log10 household income.

    Assumes that log10 income is uniform in each range.

    df: DataFrame with columns income and freq
    log_upper: log10 of the assumed upper bound for the highest range

    returns: NumPy array of log10 household income
    """
    # compute the log10 of the upper bound for each range
    df['log_upper'] = np.log10(df.income)

    # get the lower bounds by shifting the upper bound and filling in
    # the first element
    df['log_lower'] = df.log_upper.shift(1)
    df.log_lower[0] = 3.0

    # plug in a value for the unknown upper bound of the highest range
    df.log_upper[41] = log_upper

    # use the freq column to generate the right number of values in
    # each range
    arrays = []
    for _, row in df.iterrows():
        vals = np.linspace(row.log_lower, row.log_upper, int(row.freq))
        arrays.append(vals)

    # collect the arrays into a single sample
    log_sample = np.concatenate(arrays)
    return log_sample
```

In [4]:
```
def raw_moment(xs, k):
    return sum(x**k for x in xs) / len(xs)
```

In [5]:
```
def central_moment(xs , k):
    mean = raw_moment(xs, 1)
    return sum((x - mean)**k for x in xs) / len(xs)
```

In [6]:
```
def standardized_moment(xs , k):
    var = central_moment(xs, 2)
    std = math.sqrt(var)
    return central_moment(xs, k) / std**k
```

In [7]:

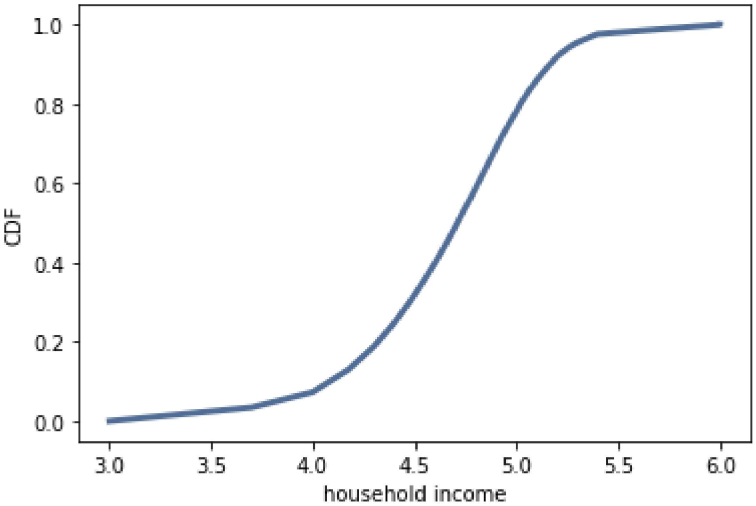
```
def skewness(xs):
    return standardized_moment(xs, 3)
```

In [8]:

```
def pearson_median_skewness(xs):
    median = np.median(xs)
    mean = xs.mean()
    var = central_moment(xs, 2)
    std = math.sqrt(var)
    return 3 * (mean - median) / std
```

In [9]:

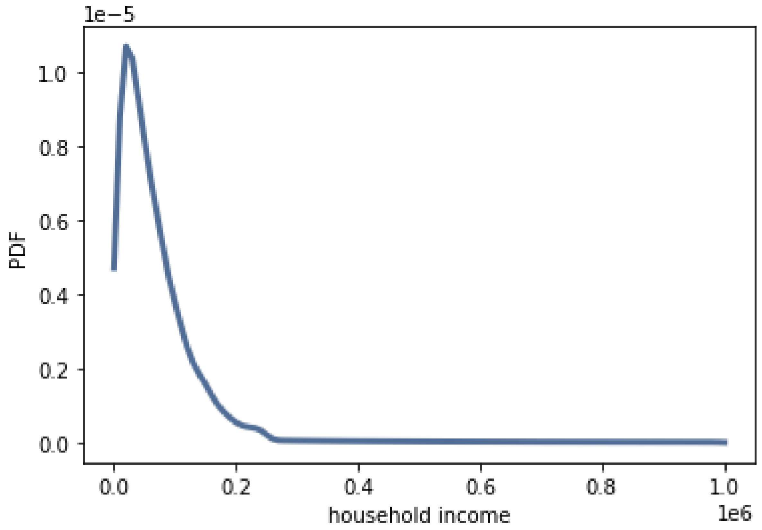
```
df = hinc.ReadData()
log_sample = InterpolateSample(df, log_upper=6)

log_cdf = thinkstats2.Cdf(log_sample)
thinkplot.Cdf(log_cdf)
thinkplot.Show(xlabel='household income',
               ylabel='CDF')

sample = np.power(10, log_sample)
```



```
<Figure size 576x432 with 0 Axes>
```

In [10]:

```
pdf = thinkstats2.EstimatedPdf(sample)
thinkplot.Pdf(pdf)
thinkplot.Show(xlabel='household income', ylabel='PDF')
```



```
<Figure size 576x432 with 0 Axes>
```

In [11]:

```
mean = round(sample.mean(), 2)
print('The mean of the sample is {}'.format(mean))

median = round(np.median(sample), 2)
print('The median of the sample is {}'.format(median))

skew = round(skewness(sample), 2)
print('The skewness of the sample is {}'.format(skew))

pm_skew = round(pearson_median_skewness(sample), 2)
print('The Pearson Median Skewness of the sample is {}'.format(pm_skew))

cdf = thinkstats2.Cdf(sample)
print('The fraction of households that report a taxable income below the mean is {}'.f

print()
print('Changing the upper bound only effects the final sample group of 2911 people wit
      'The median will not change when the upper bound is changed because the number o
      'The mean will go up or down respective to whether the upper bound is increased 
      'have a dramatic impact as only the final 2910 incomes will be affected.\n'
      'The skewness will grow or shrink depending on whether the upper bound is increa
      'The Pearson Median Skewness actually does the opposite of what is expected, it 
      'is increased.  This seems to be because the standard deviation grows faster tha
      'bound is increased.')
```

```
The mean of the sample is 74278.71
The median of the sample is 51226.93
The skewness of the sample is 4.95
The Pearson Median Skewness of the sample is 0.74
The fraction of households that report a taxable income below the mean is 0.66

Changing the upper bound only effects the final sample group of 2911 people with inco
me over $250,000.
The median will not change when the upper bound is changed because the number of peop
le has not changed.
The mean will go up or down respective to whether the upper bound is increased or dec
reased.  It will not
have a dramatic impact as only the final 2910 incomes will be affected.
The skewness will grow or shrink depending on whether the upper bound is increased or
decreased.
The Pearson Median Skewness actually does the opposite of what is expected, it shrink
s as the upper bound
is increased.  This seems to be because the standard deviation grows faster than the
mean when the upper
bound is increased.
```