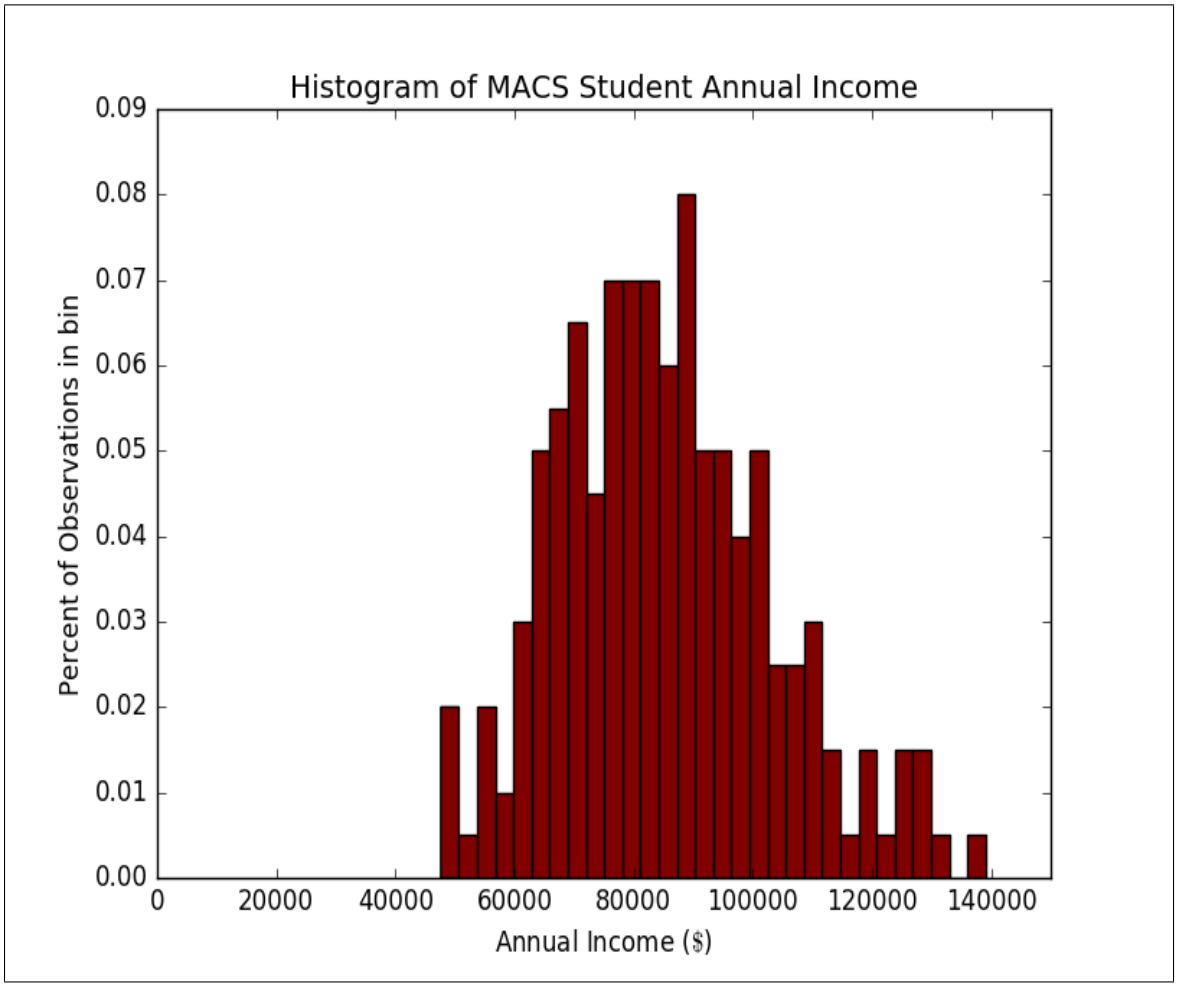
Problem Set #2

MACS 30100, Dr. Evans
Dongping Zhang

Problem 1: **Using maximum likelihood estimation (MLE) technique to estimate the parameters of a lognormal distribution that is supposedly to resemble the actual distribution of UChicago MACSS students' future incomes.**

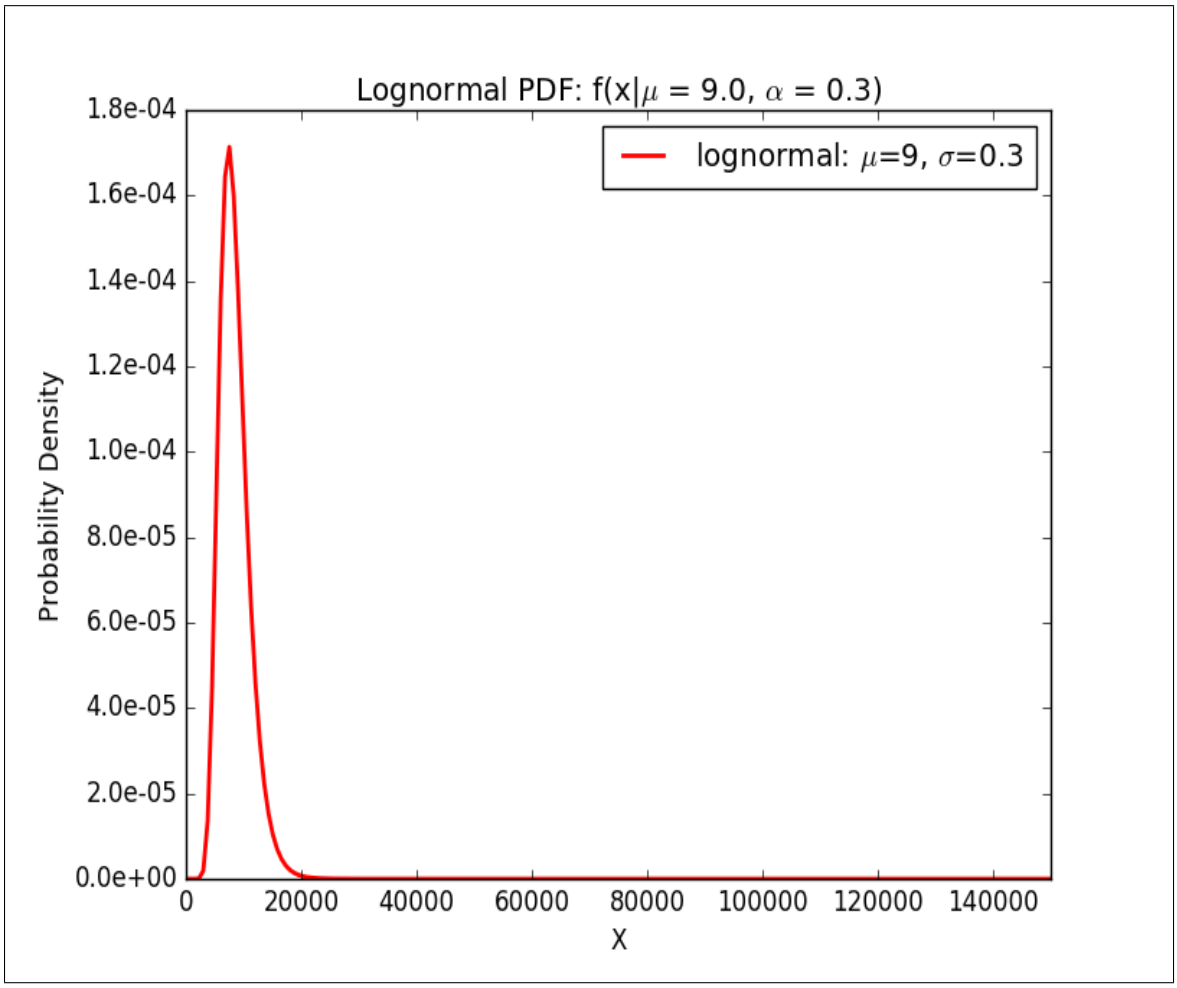
Part (a). Plot a histogram of percentages using *income.txt* data with 30 bins.

Figure 1: Probability Density Histogram of MACSS Students' Annual Income



Part (b). Plot the lognormal PDF $f(x|\mu = 9.0, \sigma = 0.3)$ for $0 \leq x \leq 150,000$.

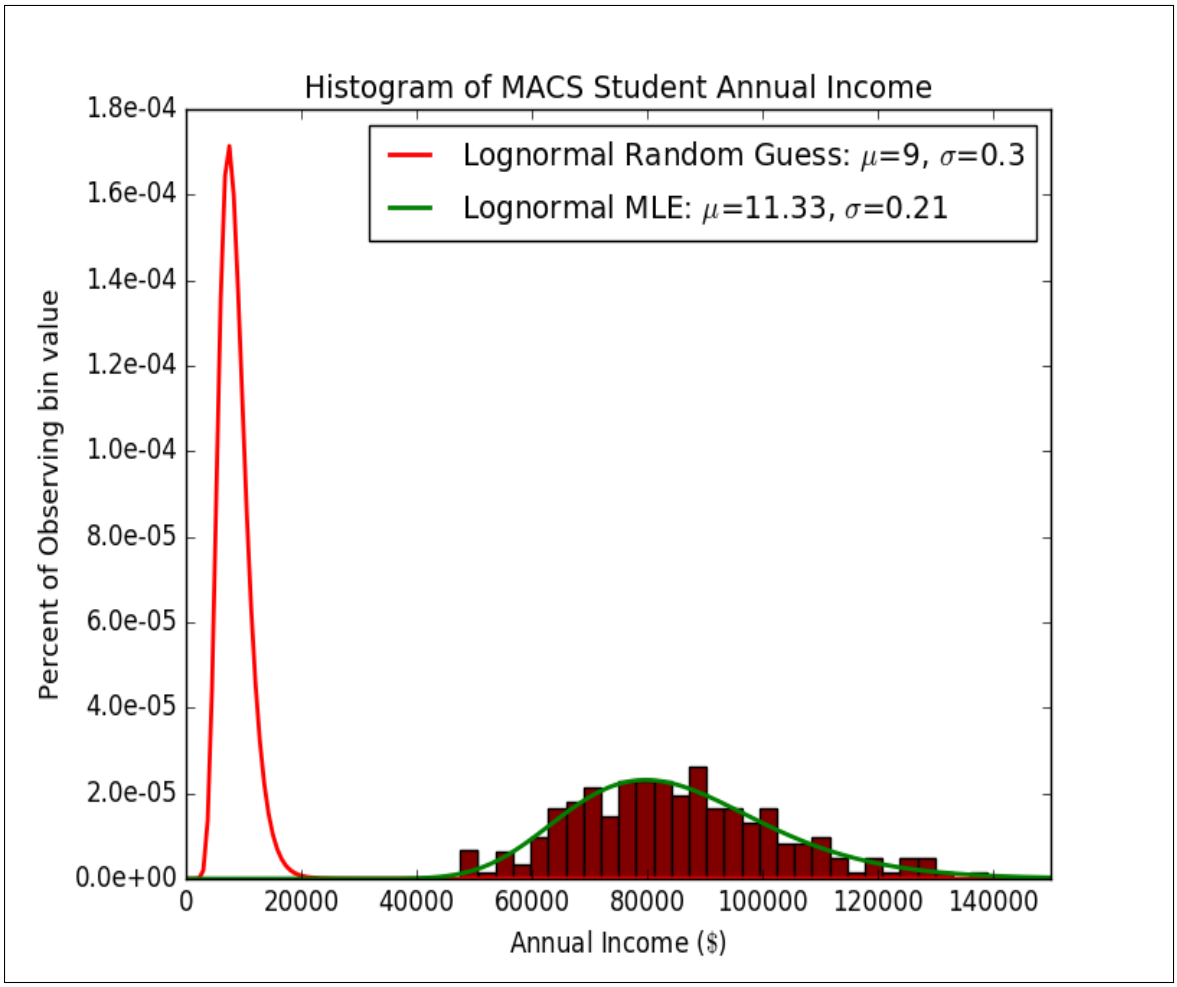
Figure 2: Lognormal PDF $f(x|\mu = 9.0, \sigma = 0.3)$



Given the income data, the log likelihood value for this parameterization is: **-8298.63695601.**

Part (c). Estimate the parameters of the lognormal distribution by maximum likelihood and plot its PDF against the PDF from part (b) and the histogram from part (a). Plot the estimated PDF for $0 \leq x \leq 150,000$. Report the ML estimates for μ and σ, the value of the likelihood function, and the variance-covariance matrix.

Figure 3: Lognormals and Income Histogram



The estimated parameters of the lognormal distribution by MLE are:

- $\mu =$ **11.3314403259**
- $\sigma =$ **0.211674582712**

The value of the likelihood function is: **-2239.534744**

The variance-covariance matrix is: $\begin{bmatrix} 2.259 \times 10^{-4} & 3.984 \times 10^{-6} \\ 3.984 \times 10^{-6} & 1.184 \times 10^{-4} \end{bmatrix}$

Part (d). Perform a likelihood ratio test to determine the probability that the data in *incomes.txt* came from the distribution in part (b).

The likelihood ratio test statistics is **12118.204424**, and the p-value for this test statistics under a chi-squared distribution with dof = 2 is: **0.0**

Thus, we would reject the H_0 and be able to claim *income.txt* does not come from the distribution of lognormal ($\mu = 9.0, \sigma = 0.3$).

Part (e). Using that estimated model from part (c), What is the probability that you will earn more than \$100,000? What is the probability that you will earn less than \$75,000?

Using the estimated model from part (c), which is a lognormal ($\mu = 11.33, \sigma = 0.21$), the estimated probability of observing a UChicago MACSS graduate to earn an income of more than \$100,000 after graduation under a lognormal distribution using MLE parameters is: 0.195617995534 or **19.56%**

Using the estimated model from part (c), which is a lognormal ($\mu = 11.33, \sigma = 0.21$), the estimated probability of observing a UChicago MACSS graduate to earn an income of less than \$75,000 after graduation under a lognormal distribution using MLE parameters is: 0.307939622515 or **30.79%**

Problem 2: **Use MLE to estimate the parameters of a linear regression based on the assumption of normally distributed error term**

The proposed sickness model is:

$$sick_i = \beta_0 + \beta_1 age_i + \beta_2 children_i + \beta_3 temp\_winter_i + \varepsilon_i \quad (1)$$
$$where \quad \varepsilon_i \sim N(0, \sigma^2)$$

which means the error term, ε_i, after conducting a linear transformation, could be expressed as

$$\varepsilon_i = sick_i - \beta_0 - \beta_1 age_i - \beta_2 children_i - \beta_3 temp\_winter_i \ \sim \ N(0, \sigma^2) \quad (2)$$

Part (a). Estimate $(\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2)$ to maximize the likelihood of seeing the data in *sick.txt*. Report your estimates, the value of the log likelihood function, and the estimated variance covariance matrix of the estimates.
After using MLE, the estimated parameters are:

- $\beta_0^{MLE} = 0.251646223604$
- $\beta_0^{MLE} = 0.0129333477208$
- $\beta_0^{MLE} = 0.400502027962$
- $\beta_0^{MLE} =$ -0.00999166717288
- $\sigma^2_{MLE} = \sigma_{MLE} \times \sigma_{MLE} = 0.0030176954401 \times 0.0030176954401 = 9.11 \times 10^{-6}$

The value generated by the log-likelihood function is: **876.865047495**

In order to obtain the variance-covariance matrix, I used the MLE estimators to implement the same optimization using 'L-BFGS-B' method. My function returned a variance-covariance matrix looks like: $\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$. This matrix does not look right to me, thus I used the same initial values, instead of the MLE parameters, and implemented the optimization procedure again using 'BFGS' method. The 'BFGS' optimization method has the same log-likelihood value and MLE parameters, and returned an inverse hessian of

$$\begin{bmatrix} 9.06 \times 10^{-7} & 1.16 \times 10^{-8} & -2.01 \times 10^{-7} & -2.29 \times 10^{-8} & 3.10 \times 10^{-9} \\ 1.16 \times 10^{-8} & 3.88 \times 10^{-9} & -3.41 \times 10^{-8} & -2.54 \times 10^{-9} & -5.60 \times 10^{-11} \\ -2.01 \times 10^{-7} & -3.41 \times 10^{-8} & 3.52 \times 10^{-7} & 2.26 \times 10^{-8} & -8.93 \times 10^{-10} \\ -2.29 \times 10^{-8} & -2.54 \times 10^{-9} & 2.26 \times 10^{-8} & 2.02 \times 10^{-9} & -2.17 \times 10^{-12} \\ 3.10 \times 10^{-9} & -5.60 \times 10^{-11} & -8.93 \times 10^{-10} & -2.17 \times 10^{-12} & 2.31 \times 10^{-8} \end{bmatrix},$$

which looked more right to me.

Part (b). Use a likelihood ratio test to determine the probability that $\beta_0 = 1.0, \sigma^2 = 0.01$ and $\beta_1, \beta_2, \beta_3 = 0$. That is, what is the likelihood that age, number of children, and average winter temperature have no effect on the number of sick days?

The likelihood ratio test statistics is: **6261.13147107**, and the p-value for this test statistics under a chi-squared distribution with dof = 5 is: **0.0**

Thus, we would reject the H_0 and be able to claim the likelihood of the statement that "age, number of children, and average winter temperature have no effect on the number of sick days" is infinitesimal.