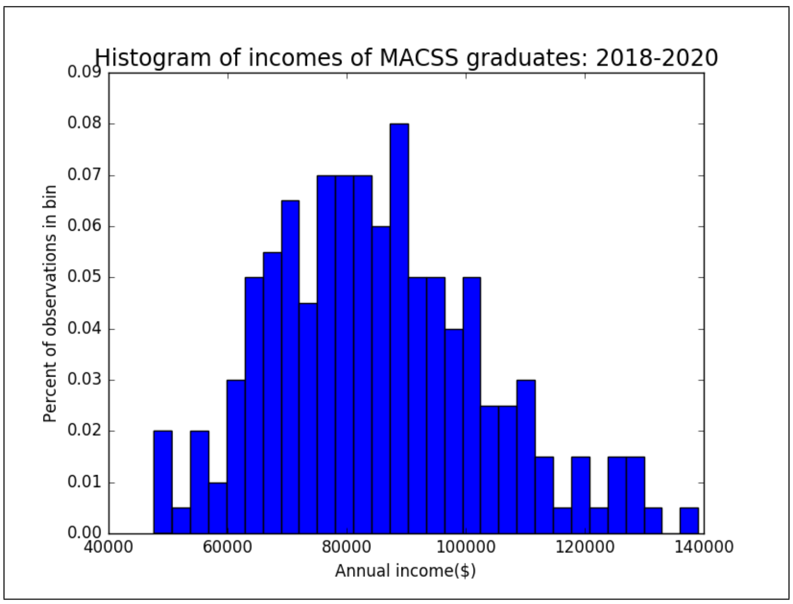
Problem Set #[2]

MACS 30100, Dr. Evans
Xuancheng Qian

1. Some income data, lognormal distribution, and hypothesis testing.
Part (a). Plot a histogram of percentages of the income.txt data with 30 bins.
Answer: Figure 1 is the histogram of percentages of the income.txt data with 30 bins.

Figure 1: Histogram of incomes of MACSS graduates: 2018-2020



Part (b). Plot the lognormal PDF and what is the value of the log likelihood value for this parameterization of the distribution and given this data?
Answer: The log likelihood value for this parameterization of the distribution and given the data is -8298.64.
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).
Answer:. By maximum likelihood estimation, we have $\mu_{MLE} = 11.33$ and $\sigma_{MLE} = 0.21$. The value of the log likelihood function is -2239.53. The variance and covariance matrix is

$$\begin{bmatrix} 1.51441374e-04 & 7.36566926e-06 \\ 7.36566926e-06 & 9.78129088e-05 \end{bmatrix}$$

Figure 2: Lognormal pdf with μ=9.0 and σ=0.3

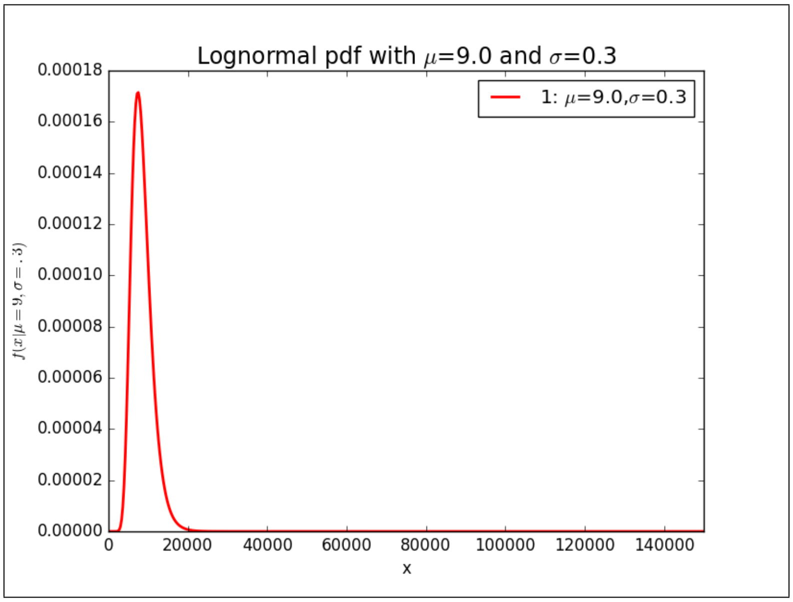
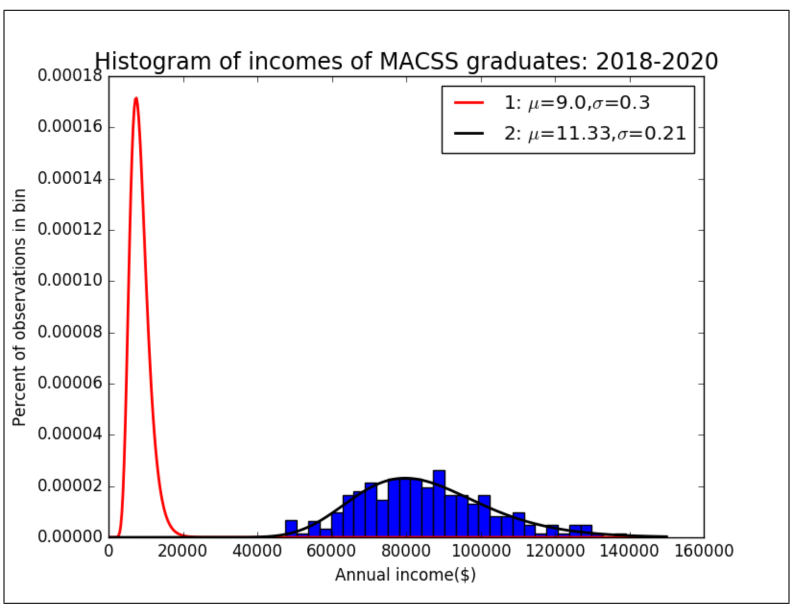


Figure 3: Histogram of incomes of MACSS graduates: 2018-2020



Part (d). Perform a likelihood ratio test to determine the probability that the data in incomes.txt came from the distribution in part (b).
Answer: The Chi square of H0 with 2 degrees of freedom p-value is 0.00 which is too small. Then we should reject the null hypothesis. This indicates that data in incomes.txt is very unlikely to come from the distribution in part(b).
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?.
Answer: The probability that I will earn more than $100,000 is 0.196. The probability that I will earn less than $75,000 is 0.308.

2. Linear regression and MLE.
Part (a). Estimate the parameters of the model by MLE. Report the estimates, the value of the log likelihood function, and the estimated variance covariance matrix of the estimates.
Answer: By maximum likelihood estimation, we have $\sigma_{MLE} = 0.0030$, $\beta_0 = 0.2516$, $beta_1 = 0.0129$, $beta_2 = 0.4005$, $beta_3 = -0.00999$. The log likelihood of function value is 876.87.
The variance and covariance matrix is

$$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Part (b). What is the likelihood that age, number of children, and average winter temperature have no effect on the number of sick days?
Answer: Likelihood ration test p-value is 0.00. So we should reject the null hypothesis and it is unlikely that age, number of children, and average winter temperature have no effect on the number of sick days.