Question 2 (a)

In this part, we first calculate the average cost for evenly distributed case, which is 0.165 for 100 points. Then we do the simulation, in which we randomly sample 100 points from the square region and use k-means method to find the centroid and calculate average cost. The simulated result is shown in Figure 1.

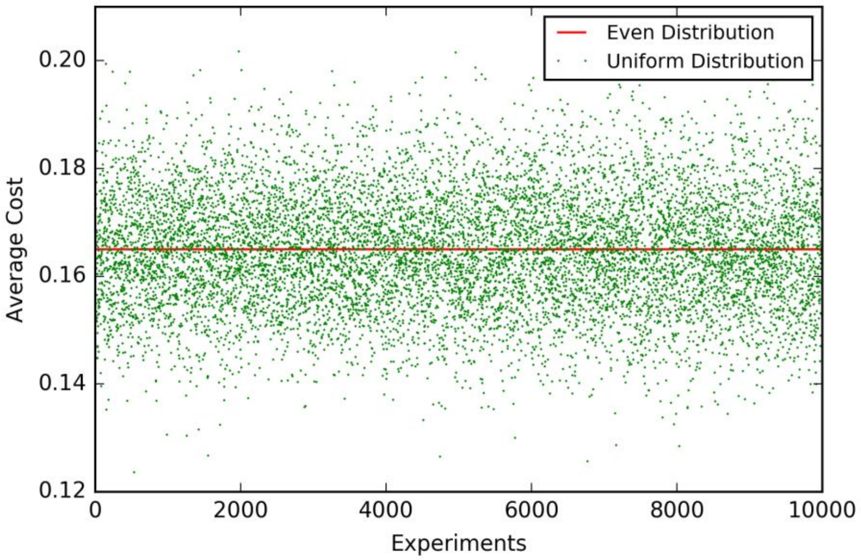


Fig. 1. Simulated result

We notice that, among those 10,000 simulated result, there are 51.44% of all experiments whose average cost is smaller than 0.165. So, we think that the randomly sampling cases are more likely to have the smaller cost.

(b)

In this part, we first simulate the cases when the number of points increases. The simulated result is shown in Figure 2. From figure 2, we can see that as the number of points increases, these two cases will have almost the same average cost.

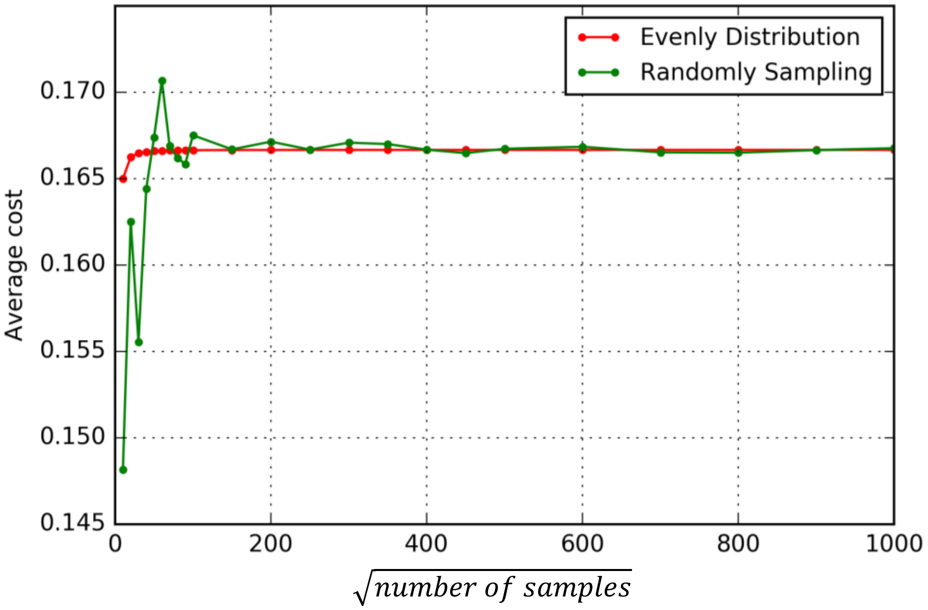
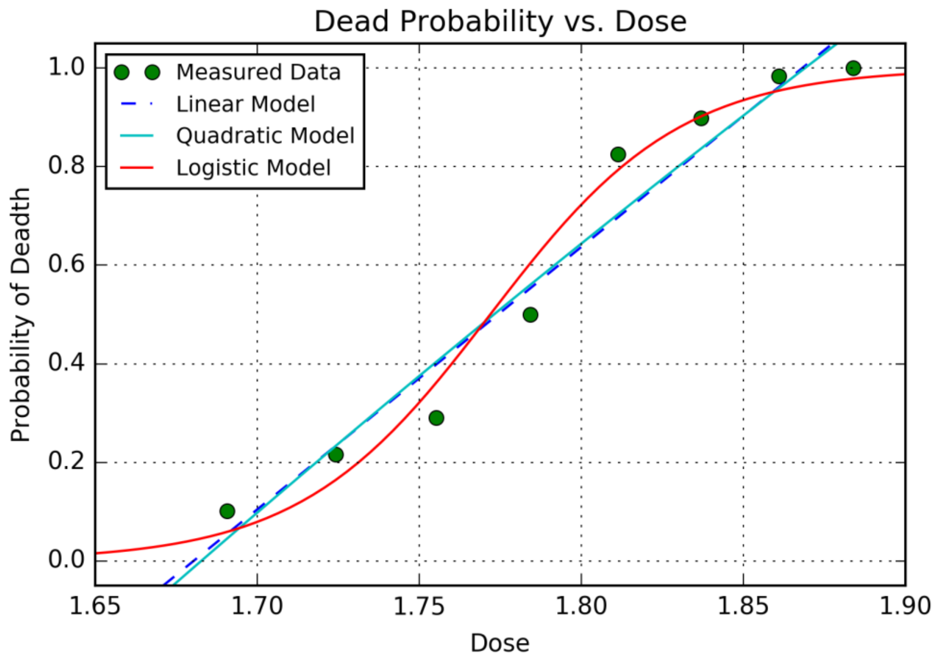


Fig. 2. Average cost vs. $\sqrt{number\ of\ samples}$

Question 5.

The fitted curves for linear, quadratic and logistic model are shown below.



Suppose the dose is represented by x and dead probability is denoted by y, then the three models are:

- Linear model: $y_1 = 5.3249\,x - 8.9478$
- Quadratic model: $y_2 = -1.7666\,x^2 + 11.6429\,x - 14.5895$
- Logistic model: $y_3 = \dfrac{1}{1+e^{-(34.1213x-60.4591)}}$

The RSS, which is used to calculate the likelihood, are:

- Linear model: $RSS_1 = 0.045091$
- Quadratic model: $RSS_2 = 0.044786$
- Logistic model: $RSS_3 = 0.022714$

Calculate AIC_c according to equation:

$$AIC_c = 2k - 2\ln(L) + \frac{2k(k+1)}{n-k-1}$$

The corresponding AIC_c for three models are:

- Linear model: 12.5982
- Quadratic model: 18.2117
- Logistic model: 13.9696

Based on above result, we should choose the linear model.