**Question 1**

Given a [dataset](http://goo.gl/cIIGzP), (file data.online.scores)which includes the records of students' exam scores for the past few years of an online course. The first column students' id, the second column is the mid-term scores, and the third column is the final scores, and data are split by tab. Based on the dataset, give out the following statistical description of data. If the result is not integer, then round it to 3 decimal places

i) (40') Give out the basic statistical description about mid-term scores.

a. (10') max = Blank 1, min = Blank 2

b. (15') First quartile Q1 = Blank 3, median = Blank 4, Third quartile Q3 = Blank 5.

c. (5') The mean score is Blank 6

d. (5') The mode score is Blank 7

e. (5') variance = Blank 8

f. (5') Do you think the model score consistent with  empirical formula that (mean - mode = 3 \* (mean - median))?  (1 for yes, 0 for no). Blank 9

**45 points**

**Question 2**

Given the inventories of two supermarkets, compare the similarity between these two supermarkets by using the different proximity measures. If the result is not integer, then round it to 3 decimal places.

a. (5')  Given 200 items, the following table summarizes how many items are supplied by corresponding supermarket:

|  |  |  |  |
| --- | --- | --- | --- |
|  | J Sainsbury | | |
| King Kullen |  | 0 | 1 |
| 0 | 31 | 18 |
| 1 | 23 | 128 |

So, the Jaccard coefficent of J Sainsbury and King Kullen is Blank 1.

b. (15') Based on all items, what's the Minkowski distance of different h values:

    1. h = 1, Minkowski Distance = Blank 2

    2. h = 2, Minkowski Distance = Blank 3

    3. h = infinite, Minkowski Distance = Blank 4

c. (5') Cosine similarity = Blank 5

d. (5') Kullback–Leibler divergence. we denote there are i\_1 of item1 in  J Sainsbury, and j\_1 of item1 in King Kullen. Assume there is a customer, who will pick up a product by random, so the probability of this customer to pick up item 1 in J Sainsbury is i\_1 / (i\_1 + ... + i\_100). Based on this probability distribution, calculate the Kullback–Leibler divergence of these two supermarkets P(J Sainsbury || King Kullen) = Blank 6

**30 points**

**Question 3**

The following table is a summary about customers' purchase history of diapers and beer. If the result is not integer, then round it to 3 decimal places

|  |  |  |
| --- | --- | --- |
|  | by diaper | not buy diaper |
| buy beer | 1976 | 283 |
| not buy beer | 128 | 83712 |

Calculate the chi-square correlation value.

Answer

**10 points**

**Question 4**

Based on the students' scores data, give out the following correlation analysis: (if the result is not integer, then round it to 3 decimal places)

a. Pearson's correlation coefficient between mid-term scores and final scores =  Blank 1

b. Covariance between mid-term scores and final scores  = Blank 2

**10 points**

**Question 5**

Based on the data of students' score, please normalize the mid-term score using z-score normalization. What's the new score of 90 after normalization?

If the result is not integer, then round it to 3 decimal places

Answer

**10 points**