Sam Schlesinger
CS412_Assignment1

Problem 1

a. The maximum of any set can be found by doing element-wise inequality comparison over the set and storing the greatest value. Likewise, the minimum can be found by doing element-wise inequality comparison and storing the least value.
For the given dataset, Max = 100, Min = 37.

b. Median of any set can be computed by sorting the set and taking the *middle value* which separates the higher half of the dataset from the lower half. In the case of an odd number of data elements, this value will be unique. In our case, we have an even number of data elements, so the median is not unique and is calculated by convention as the average of the two middle elements, ie, for a set of size n,
median = $(x_{n/2} + x_{(n/2)+1})/2$ Likewise, the first quartile and the third quartile can be computed by finding the median of the lower half and higher half of the data respectively.
For the given dataset, First Quartile Q1 = 68, Median = 77, Third Quartile Q3 = 87

c. For any set of n numbers X = $\{x_1 ... x_n\}$, the mean can be computed as $\mu = 1/n \sum_{i=1}^{n} x_i$.
For the given dataset, mean μ = 76.715.

d. The mode of a set can be found by counting the frequency of each discrete value of a set. The value with the most instances is the mode. For the given dataset we actually have two values of equal frequency, thus the dataset can be considered multimodal.
mode = [77, 83]

e. Variance σ^2 of a set of N observations is calculated as $\sigma^2 = 1/N \sum_{i=1}^{N} (x_i - \bar{x})^2$ where $\bar{x}$ is the mean value of the observations.
For the given dataset, Variance σ^2 = 173.106

Problem 2

The median of a histogram can be approximated by interpolation using the equation

$$median \approx L_m + \left(\frac{n/2 - F_{m-1}}{f_m}\right) \times (L_{m+1} - L_m)$$

The term $(L_{m+1} - L_m)$ is the width of the intervals of the histogram.

L_m is the lower boundary of the median interval, which can be obtained by summing the frequencies in each bin until adding the next bin would put the sum over n/2.

F_{m-1} is the sum of the frequencies below the median interval.

f_m is the frequency of the median interval

n is the number of observations in the set and can be calculated by adding the frequency of each bin

For this problem, the variables are as follows:

n = 493; $(L_{m+1} - L_m)$ =5; L_m=25; F_{m-1} = 189; f_m = 92

Plugging into the equation, we get:

$$\underline{median} \approx 25 + \left(\frac{493/2 - 189}{92}\right) \times 5 = \underline{28.12}$$

Problem 3

Normalized midterm scores are computed using the z-score equation $z = (x-\mu)/\sigma$ for each element x in the dataset where μ = mean of the population and σ = standard deviation of the population.

a. Variance of midterm-original, as computed in problem 1e, is $\sigma^2 = 177.106$. Variance of midterm-normalized is $\sigma^2 = 1.0$ as computed using the variance equation from 1e and substituting the new scores z_i and the new mean $\bar{z}$ (for x_i and $\bar{x}$) after performing z-score normalization described above. This is expected functionality as the purpose of z-score normalization is to center the data around 0 with variance of 1.

b. Using the z-score equation laid out above with:
μ = 76.715 (as calculated in 1c)
σ = sqrt(Variance) = 13.308 (square root of the variance calculated in 1e)
Z_{90} = (90-76.715)/13.308; $Z_{90} = 1.009$

c. The Pearson's correlation coefficient of two attributes A and B of a set of size n can be calculated using the equation:

$$r_{A,B} = \frac{\sum_{i=1}^{n} (a_i - \bar{A})(b_i - \bar{B})}{n\sigma_A \sigma_B}$$

Where $\bar{A}$, $\bar{B}$and σ_A, $\sigma_B$are the Expected Value (mean) and standard deviation of sets A and B, respectively. Per 1c and 3b, $\bar{A}$ and $\sigma_A$are 76.715 and 13.308 respectively. Solving for $\bar{B}$ and $\sigma_B$using the methodology laid out in 1c and 1e/3b, we find that $\bar{B}$= 87.084 and σ_B=10.914. Solving for $r_{A,B} = 0.544$

d. Using the methodology laid out in 3c above and substituting our midterm-normalized data ($\bar{A}$= 0 and σ_A= 1), we get $r_{A,B} = 0.544$. Notice that this is the same value as 3c before normalization. This is because the two attributes are still correlated to the same degree as A has simply been normalized around 0 which is a linear transformation.

e. Knowing the relationship between Covariance and Correlation Coefficient: $r_{A,B} = Cov(A,B)/\sigma_A\sigma_B$, we can solve for Cov(A,B) simply by multiplying our correlation coefficient by $\sigma_A\sigma_B$. Doing this we get, Cov(A,B) = 78.176

Problem 4

a. Minkowski difference of two l-dimensional vectors i and j (where i=(x_{i1},x_{i2},...x_{il}) and j=(x_{j1},x_{j2},...x_{jl}))can be computed as:

$$d(i,j) = \sqrt[p]{|x_{i1} - x_{j1}|^{p} + |x_{i2} - x_{j2}|^{p} + \ldots + |x_{il} - x_{jl}|^{p}}$$

Plugging in different values of h and solving:

i. d(i,j,h=1) = 6152.0
ii. d(i,j,h=2) = 715.328
iii. d(i,j,h=∞) = 170.0

b. Cosine similarity of two vectors d_1 and d_2 is calculated using the equation:

$$cos(d_1,d_2) = \frac{d_1 \bullet d_2}{||d_1|| \times ||d_2||}$$

Where $d_1 \bullet d_2$ is the vector dot product of the two vectors and ||d|| is the length of the vector. Plugging in our values for CML and CBL, we get, cos(CML, CBL) = 0.841.

c. Kullback Liebler divergence D_{KL} between two probability distributions p(x) and q(x) in it's discrete form can be calculated using the equation:

$$D_{KL}(p(x)||q(x)) = \sum_{x \in X} p(x)\, ln\frac{p(x)}{q(x)} dx$$

As described in the problem statement, we first compute our probability distributions for each library by dividing each entry for the number of an individual book in the library by the total number of books in the library, i.e. p(Book 1) = $\frac{i_1}{i_1 + \ldots + i_{100}}$. We then calculate D_{KL}(CML||CBL) by substituting p(CML) for p(x) and p(CBL) for q(x). It is also worth noting here that D_{KL}(CML||CBL) is not the same as D_{KL}(CBL||CML), so the ordering of the substitution is important. Performing the substitution yields: D_{KL}(CML||CBL) = 0.207

Problem 5

a. The distance measure for symmetric binary variables i and j can be calculated as:

d(i,j) = $\frac{r+s}{q+r+s+t}$ given the binary contingency table:

	Object j		
Object i		1	0
	1	q	r
	0	s	t

Substituting the values from the table given in the problem statement, we get:

d(Buy Beer,Buy Diaper) = $\frac{40+15}{150+40+15+3300} = \frac{55}{3505} \approx 0.0157$

b. Jaccard coefficient for the similarity between two asymmetric binary variables i and j can be calculated using the equation:

$sim_{Jaccard}(i,j) = \frac{q}{q+r+s}$ given the same table as laid out in 5a

Substituting from the table given in the problem statement, we get:

$sim_{Jaccard}$(Buy Beer,Buy Diaper) = $\frac{150}{150+40+15} = \frac{150}{205} \approx 0.732$

c. The chi-squared statistic for a contingency table with nominal attributes A and B of dimension c x r can be calculated using the formula:

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Where o_{ij} is the observed value and e_{ij} is the expected frequency calculated as

$e_{ij} = \frac{count(A=a_i) \times count(B=b_j)}{n}$

Substituting the values from the given table:

$e_{11} = \frac{165 \times 190}{3505} = 8.944$; $e_{10} = \frac{190 \times 3340}{3505} = 181.056$;

$e_{01} = \frac{165 \times 3315}{3505} = 156.056$; $e_{00} = \frac{3340 \times 3315}{3505} = 3158.944$

And plugging into the chi-squared equation, we get:

$\chi^2 = \frac{(150-8.944)^2}{8.944} + \frac{(40-181.056)^2}{181.056} + \frac{(15-156.056)^2}{156.056} + \frac{(3300-3158.944)^2}{3158.944}$; $\chi^2 = 2468.286$

d. At a significance level of α = 0.05, for 1 degree of freedom (calcuated as dof = (r-1)(c-1) = (2-1)(2-1)), we need a χ^2 value of 3.841 to reject the null hypothesis. The value calculated in 5c of χ^2 = 2468.286 > 3.841, therefore we must reject the null hypothesis that Buy Beer and Buy Diaper are independent. Given that the hypothesis is rejected, we then say that Buy Beer and Buy Diaper are statistically correlated.