EXAM #1

CS412, Data Mining

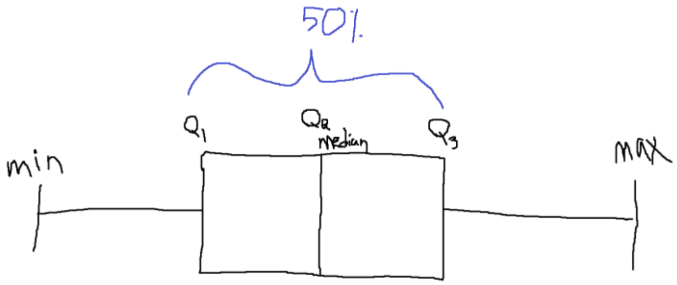
Pouya Akbarzadeh (pa2)

March 22, 2021

Question 1

a) A five-number summary of a distribution allows us to get an overall look of our data. It allows us to see the min, Q1, Q2, Q3, and the max (Q2 is the median). Q1 is the lower quartile and Q3 is the upper quartile. All this info allows us to get a good idea of what the distribution is at a quick glance. However, a lot of info about our data is lost through summarizing

b) Boxplot is a graphical method of displaying data using quartiles. Boxplots are used to show the five-number summary of the data (explained in part a). Allow me to draw it out below.



c) Yes, they can. However, just because the boxplot or the five-number summary is the same, does not mean that the distribution is the same. This can be easily explained when we are unable to understand the distribution of the data between the Min-Q1, Q1-Q2, Q2-Q3, and Q3-Max. This means that we do not know what is going on between those values. So, 2 different sets of data can yield the same five-number summary.

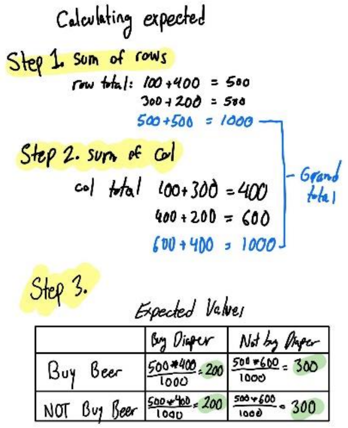
d) Quantile is where we take our data and divide into equal-sized continuous intervals. A quantile plot allows us to see the shape of our data's distribution by graphically representing the quantiles.

e) Quantile-Quantile plots are two quantile plots (explained in part d) compared against each other. Quantile-Quantile plots are used to make continuous distribution comparisons.

f) As explained in parts e and d, quantile plots display the quantiles of our data. However, quantile-quantile plots allow us to compare 2 different quantiles of two sets of numbers. Both display the values corresponding to their data's distribution; however, we add a line and divide it up based on Q1, Q2, and Q3. This allows more info being communicated from the plot. Points that are above this y=x line show a higher value for the distribution, and vice versa.

Question 2

a. 200 (See calculation below)
b. 300 (See calculation below)



Calculating expected

Step 1. sum of rows

row total: 100 + 400 = 500
300 + 200 = 500
500 + 500 = 1000

Step 2. sum of Col

col total 100 + 300 = 400
400 + 200 = 600
600 + 400 = 1000

- Grand total

Step 3.

Expected Values

	Buy Diaper	Not by Diaper
Buy Beer	$\frac{500*400}{1000}$ = 200	$\frac{500*600}{1000}$ = 300
NOT Buy Beer	$\frac{500*400}{1000}$ = 200	$\frac{500*600}{1000}$ = 300

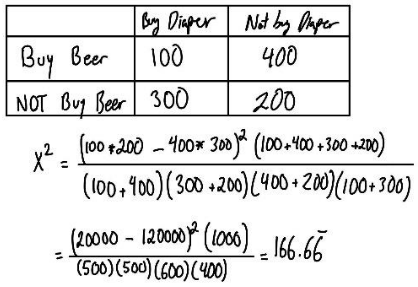
Row Total:
100 + 400 = 500
300 + 200 = 500

Col Total :
100 + 300 = 400
400 + 200 = 600

Grand total = 1000

(500 * 400) / 1000 = 200
(500 * 400) / 1000 = 200
(500 * 600) / 1000 = 300
(500 * 600) / 1000 = 300

c. 166.666



	Buy Diaper	Not by Diaper
Buy Beer	100	400
NOT Buy Beer	300	200

$$X^2 = \frac{(100*200 - 400*300)^2 (100+400+300+200)}{(100+400)(300+200)(400+200)(100+300)}$$

$$= \frac{(20000 - 120000)^2 (1000)}{(500)(500)(600)(400)} = 166.6\bar{6}$$

$$166.666 = \frac{(100*200 - 400*300)^2(100+400+200)}{(100+400)\ (300+200)\ (400+200)(100+300)}$$

d. We look at the significance level and not that it is 0.05. At this value and considering independence and considering our p value (<0.05) we can state that ‘Buy Beer’ and ‘Buy Diaper’ are independent. Our p value is 1/10000 which is clearly less than 0.05.

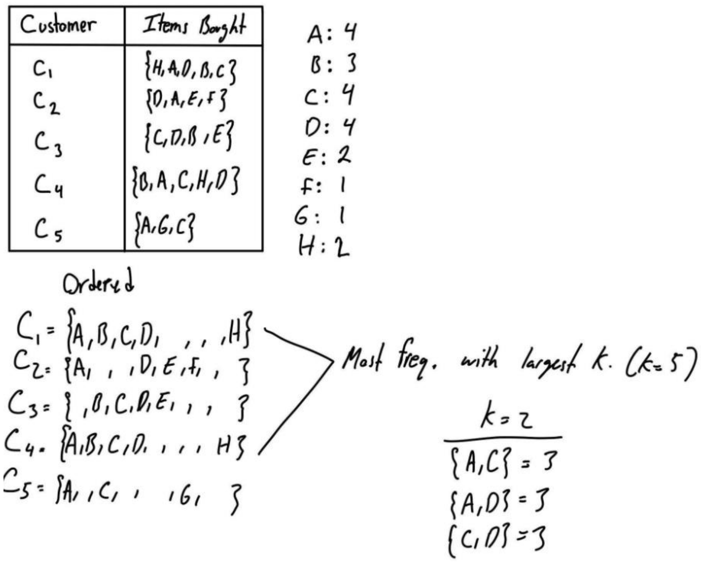
e. 887.6113204

$$887.6113204 = \frac{(100*20000 - 400*300)^2\ (100+400+20000)}{(100+400)\ (300\ +\ 20000)\ (400+20000)(100+300)}$$

f. We look at the significance level and not that it is 0.05. At this value and considering independence and considering our p value (<0.05) we can state that ‘Buy Beer’ and ‘Buy Diaper’ are independent. Our p value is 1/10000 which is clearly less than 0.05. p-value was calculated using: https://www.socscistatistics.com/pvalues/chidistribution.aspx

Question 3

a.



Customer	Items Bought
C_1	{H,A,D,B,C}
C_2	{D,A,E,F}
C_3	{C,D,B,E}
C_4	{B,A,C,H,D}
C_5	{A,G,C}

A: 4
B: 3
C: 4
D: 4
E: 2
F: 1
G: 1
H: 2

Ordered

C_1 = {A,B,C,D, , , ,H}
C_2= {A, , ,D,E,F, , }
C_3= { ,B,C,D,E, , , }
C_4. {A,B,C,D, , , , H}
C_5= {A, ,C, , ,G, }

Most freq. with largest k. (k=5)

k=2
{A,C} = 3
{A,D} = 3
{C,D} = 3

i. I believe the largest k = 5. The 2 transactions that contain 5 items are C1 and C4 (as seen above).

ii. Having 5 transactions and having requirement of min support of 0.6 means that 3 or more of them must have the relation for our statement. Thus we are left with only 1 statement: (B ∧ C) → D [s=0.6, c=1]. With support being 3/5 = 0,6 and confidence of 3/3.

b.

i. As seen in class, the first step is to order the value of each item in ascending order. Once the constraint of the average price is achieved any additional items added to the pattern will meet the condition. They will satisfy the condition because the value of the item will be higher than our average. Thus, it is monotonic. And since we ordered the values, we can say that it is also convertible. Thus, it is both convertible and monotonic.

ii. I believe because this constraint states that there will be at least a $5 profit margin and each pattern is at least $200 that it is a rule. We can see that it is succinct, where "Succinctness: If the constraint c can be enforced by directly manipulating the data". Because we can simply not include data that will not allow our profit to be at least $5, thus manipulating the data. Furthermore, it is also anti-monotonic because if the profit doesn't reach that $5 mark it can be tossed.

Referenced for this question: https://compass2g.illinois.edu/bbcswebdav/pid-5266987-dt-content-rid-70159759_1/courses/cs_412_120211_199012/06FPAdvancedupdated%20%28March%201%29.pdf

Question 4

The table below is formed by the following steps. After finding all the frequent items in sequences, the length was found (length -1). They are as follows:

<a>:4	<b>:4	<c>:4	<d>:3	<e>:3	<f>:3

Where the numbers represent the support count. Partitioning takes place. We divide into the 6 prefixes a, b, c, d, e, and f. Then we find the subset of sequential patterns.

Looking at the projected database column we can see that some letters are removed. This is because we remove the specific letter that began as a prefix of the sequence. Then we remove that letter and only look at what's left. Then we move onto the sequential patterns now that we have the projected database for each letter/prefixes

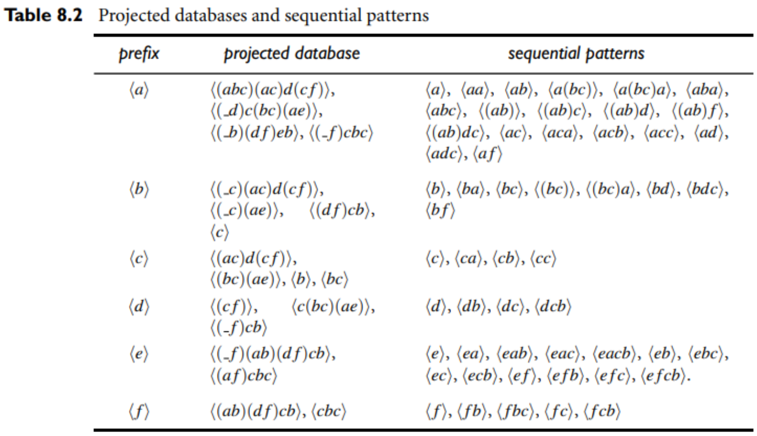


Table 8.2 Projected databases and sequential patterns

prefix	*projected database*	*sequential patterns*
$\langle a\rangle$	$\langle(abc)(ac)d(cf)\rangle$, $\langle(\_d)c(bc)(ae)\rangle$, $\langle(\_b)(df)eb\rangle$, $\langle(\_f)cbc\rangle$	$\langle a\rangle$, $\langle aa\rangle$, $\langle ab\rangle$, $\langle a(bc)\rangle$, $\langle a(bc)a\rangle$, $\langle aba\rangle$, $\langle abc\rangle$, $\langle(ab)\rangle$, $\langle(ab)c\rangle$, $\langle(ab)d\rangle$, $\langle(ab)f\rangle$, $\langle(ab)dc\rangle$, $\langle ac\rangle$, $\langle aca\rangle$, $\langle acb\rangle$, $\langle acc\rangle$, $\langle ad\rangle$, $\langle adc\rangle$, $\langle af\rangle$
$\langle b\rangle$	$\langle(\_c)(ac)d(cf)\rangle$, $\langle(\_c)(ae)\rangle$, $\langle(df)cb\rangle$, $\langle c\rangle$	$\langle b\rangle$, $\langle ba\rangle$, $\langle bc\rangle$, $\langle(bc)\rangle$, $\langle(bc)a\rangle$, $\langle bd\rangle$, $\langle bdc\rangle$, $\langle bf\rangle$
$\langle c\rangle$	$\langle(ac)d(cf)\rangle$, $\langle(bc)(ae)\rangle$, $\langle b\rangle$, $\langle bc\rangle$	$\langle c\rangle$, $\langle ca\rangle$, $\langle cb\rangle$, $\langle cc\rangle$
$\langle d\rangle$	$\langle(cf)\rangle$, $\langle c(bc)(ae)\rangle$, $\langle(\_f)cb\rangle$	$\langle d\rangle$, $\langle db\rangle$, $\langle dc\rangle$, $\langle dcb\rangle$
$\langle e\rangle$	$\langle(\_f)(ab)(df)cb\rangle$, $\langle(af)cbc\rangle$	$\langle e\rangle$, $\langle ea\rangle$, $\langle eab\rangle$, $\langle eac\rangle$, $\langle eacb\rangle$, $\langle eb\rangle$, $\langle ebc\rangle$, $\langle ec\rangle$, $\langle ecb\rangle$, $\langle ef\rangle$, $\langle efb\rangle$, $\langle efc\rangle$, $\langle efcb\rangle$.
$\langle f\rangle$	$\langle(ab)(df)cb\rangle$, $\langle cbc\rangle$	$\langle f\rangle$, $\langle fb\rangle$, $\langle fbc\rangle$, $\langle fc\rangle$, $\langle fcb\rangle$

However, in this example the min support is 2, in our case it is 3. So the following will remain

Prefix	Sequential Patterns
<a>	<a>, <ab>, <a(bc)>, <ac>, <acb>, <acc>, <aa>
<b>	<b>, <bc> <bf>
<c>	<c>, <cb>, <cc>
<d>	<d>, <dc>
<e>	<e>
<f>	<f>, <fb>

We observe that e and f appear only once. This is because there are only 2 possible suffixes for them. Thus, the minimum support would be less than 3. (While in the table 8.2 min support is 2)

To go into a bit more detail we can look at prefix d. We first look at projected database containing the letter D. Since we are doing the prefix projection, we look through the data and we see that letter D appears 3 times in all the sequences in the transaction db. So we then partition the search space and then filter those sequences where A is first. We are then left with the following: <d> and <dc>

I referenced this: http://web.cs.ucla.edu/~yzsun/classes/2018Fall_CS145/Slides/chapter_8_sequence.pdf and our text book for this question.