Middle East Technical University Department of Computer Engineering

Data Mining
Homework 1

Alper KOCAMAN
2169589

Answers

1)

a) Nominal

b) Ordinal

c) Ratio

d) Ordinal

e) Nominal

f) Ratio

g) Ordinal

h) Nominal

i) Interval

2) All items have equal chance to be selected in random sampling. There is no need to break sample data into subgroups and take further stages on these groups. Sampled random numbers distributed homogeneously in lots of tests. Actual results can be reached by applying more tests with random numbers.

3)

a) Dependent variable: Catching a cold
Independent variable: Regular 30 minute workout

b) Dependent variable: Making more rational decisions
Independent variable: Meditation

4) A .png file has been added for this question in submission since I couldn't add a photo in latex.

5)

a) Mode

b) Median

c) Median

d) Mode

e) Mean

6) No mode.

7) Sorted data => $[17, 17, 17, 18, 19, 19, 26, 28, 52, 59]$
Median of the data => 19
Mean => 27.2
Mode => 17
Midrange => $(59 + 17)/2 = 38$

8) Sample variance formula and standard deviation formula needed for this question.

Sample variance formula:

$$s^2 = \frac{\displaystyle\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}$$

Sample standard deviation formula

$$s = \sqrt{\frac{\displaystyle\sum_{i=1}^{n}(x_i - \mu)^2}{n-1}}$$

Sorted elements are $3, 7, 8, 9, 10, 14$.

Mean of these elements are $\frac{3+7+8+9+10+14}{6}\ =\ 8.5$

Subtracting all elements from mean and square these differences, 65.5 is obtained.

By using above formulas,
Sample Variance $65.5/5\ =\ 13.1$.
Sample standard deviation is square root of sample variance 3.6193922141708.

9) Variance formula and standard deviation formula needed for this question.

Varience formula:

$$\sigma^2 = \frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}$$

Standard deviation formula:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

Range is the difference between maximum value and minimum value.
Maximum element is 316 and minimum element is 10.
Thus, range $= 316 - 10 = 306$

Mean of these elements are 105.3.

By appliying formulas above,

Variance is 6638.5555555 and standard deviation is 81.477331544151.

10.

If 2 events are mutually exclusive, these events can't happen at the same time.
For example, tossing a coin is a mutually exclusive event since both results can't happen for the same tossing.

If 2 events are independent, an occurrence of one event does not affect the other event.

Thus, mutually exclusive events are not independent since these events affect each other.

11.

Normal distribution formula for the random variable Y:

$$Y = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

where x is a normal random variable, μ is the mean and σ is the standard deviation.

Some characteristic properties of normal distribution:
-Symmetric
-unimodal
-asymptotic
-mean, median, and mode are all equal
-perfectly symmetrical around its center

Normal distribution graph depends on two factors:

1)Mean which determines center location
2)standard deviation which determines height and width of the graph.

12.

Standard normal distribution is a specific type of normal distribution.

Normal distribution's specifications determined by 2 factors which are mean and standard deviation as explained in the above question.

Mean is 0 and variance is 1 in standard normal distribution.
Also, the total area under the normal curve is equal to 1.

13.

Z-test => if variance is known and the sample size is large, this test used to determine whether two population means are different than the other.

T-test => if there is a relevance between 2 data group in terms of some features and these groups are wanted to be investigated by looking if there is a significant difference between their means, t test should be used. This test is mostly used when the data set follows a normal distribution and may have unknown variances.

Chi-square test => compares expectations to actual data or model results and measures difference between them.

14.

F test is used for comparing two variances or standard deviations.

15.

If 2 variables are in positive relationship, these 2 variables are increase or decrease together.

If 2 variables are in negative relationship, when one of the variables are in trend of increase,the other one will decrease.

16.

Explained Variation→ The sum of the squared of the differences between each predicted value and the mean of these values.

Unexplained variation→ The sum of the squared of the differences between the value of each ordered pair and each corresponding predicted value.

Total variation→ Sum of the squares of the differences between the value of each ordered pair and the mean of these values.

17.

Relation strength between two variables are stated by using correlation coefficients. If there is no relationship between 2 variables, correlation coefficient between variables will be 0.

18.

By adding all data values and dividing this addition to the number of data, mean is found. If a value is changed, addition will be changed as well and mean will be changed as well.

Median is the medium value when all data is sorted. If we change a value, median doesn't have to be changed.

Mode is the element which repeated at most. If the most repeated is not changed when an element changed, mode will not be change.

Only mean have to be changed.

19.

a) Standard deviation Formula is:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{n}}$$

False.
Since sample data and mean differences are squared, inside of the formula cannot be negative. Sample data size cannot be negative as well. Division of 2 positive numbers is positive as well.

b) False.
Outliers can change the mean value but change in the mean is much less.
Then, $(oulier\ data\ -\ \mu)^2$ value can greatly change the standard deviation.

c) True.
The standard normal curve is symmetric about 0 and the total area under it is 1.

d) True.
This variable skewness can be detected and "v" is right skewed.

20.

MAE (Mean Absolute Error) formula is:

$$\frac{1}{n}\Sigma_{i=1}^{n}|x_i - x|.$$

where n is the number of errors, x_i is measured value and x is true value.

$|x_i - x|$ values are : $20, 6, 3, 7, 6$.
$\Sigma_{i=1}^{n}|x_i - x| = 20 + 6 + 3 + 7 + 6\ =\ 42$.
$42/5 = 8.4$ It is found that MAE for the given linear model as 8.4.