ASSIGNMENT 1 PART 1: DISCRETE DATA

Submitted By

Zajiba Sadia Islam

Student ID: 501279357

Group-09

Submitted for

Program: MEng in Biomedical Engineering

Course: Stats for the Health Sciences

BP8101, Winter 2024

Class Timing: Wednesday 3:00-6:00 PM

SUMMARY

This assignment deals with the various statistical measurements calculated for a discrete variable dataset. In this assignment, the number of mock tests taken by 30 IELTS candidates in a certain IELTS preparatory center has been taken and various interpretations of the data has been made. The sample has been chosen randomly, that is, the candidates have been chosen at random among all the candidates of the center. The data has been represented by means of relevant graphs including histogram, dot plot, box and whiskers plot, pie chart and others, and important inferences such as frequency, mean, standard deviation, quartiles, percentiles etc. has been calculated and shown with appropriate formula and calculations. The relevant calculations have also been presented in relevant charts.

The data used here has been collected from the website [www.kaggle.com](http://www.kaggle.com)

The data of the number of mock tests taken by 30 different IELTS candidates have been taken. It is presented in the following graph with the serial number of the candidate and the number of mock tests taken by each one of them.

|  |  |
| --- | --- |
| Serial no. of candidate | Number of mock tests taken |
| 1 | 5 |
| 2 | 3 |
| 3 | 8 |
| 4 | 4 |
| 5 | 6 |
| 6 | 5 |
| 7 | 3 |
| 8 | 7 |
| 9 | 4 |
| 10 | 4 |
| 11 | 6 |
| 12 | 5 |
| 13 | 3 |
| 14 | 8 |
| 15 | 4 |
| 16 | 6 |
| 17 | 5 |
| 18 | 3 |
| 19 | 7 |
| 20 | 4 |
| 21 | 4 |
| 22 | 6 |
| 23 | 5 |
| 24 | 6 |
| 25 | 8 |
| 26 | 7 |
| 27 | 3 |
| 28 | 0 |
| 29 | 4 |
| 30 | 6 |

Here, we see that the total number of candidates, n= 30

Maximum number of tests taken is = 8

The minimum number of tests taken is = 0

From these values, we construct the following frequency table

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Number of mock tests taken | Frequency of taking the test (f) | Relative frequency (f/n) | Cumulative relative frequency | Percentage  (relative frequency\*100) |
| 0 | 1 | 0.03 | 0.03 | 3 |
| 1 | 0 | 0 | 0.03 | 0 |
| 2 | 0 | 0 | 0.03 | 0 |
| 3 | 5 | 0.17 | 0.2 | 17 |
| 4 | 7 | 0.23 | 0.43 | 23 |
| 5 | 5 | 0.17 | 0.6 | 17 |
| 6 | 6 | 0.2 | 0.8 | 20 |
| 7 | 3 | 0.1 | 0.9 | 10 |
| 8 | 3 | 0.1 | 1 | 10 |

Here, the number of mock tests taken by IELTS candidates to prepare for the exam is taken as the dataset. The population would be the data of the number of mock tests taken by all the IELTS candidates in the centre. However, the data of 30 candidates is taken at random. Hence, the sampling technique used is simple random sample.

The data presented here is discrete. This is because the number of tests that can be taken is of certain definite integer value. The values are finite and countable. Hence, the dataset is discrete.

The data of the mock tests taken is collected in the timeframe of 6 months before the exam. Hence, the number of events (mock tests) happened during a fixed time (within 6 months before the final exam). Moreover, the events are random and independent of one another. Taking all these characteristics in consideration, it can be said that the data follows Poisson Probability Distribution.

Graphs:

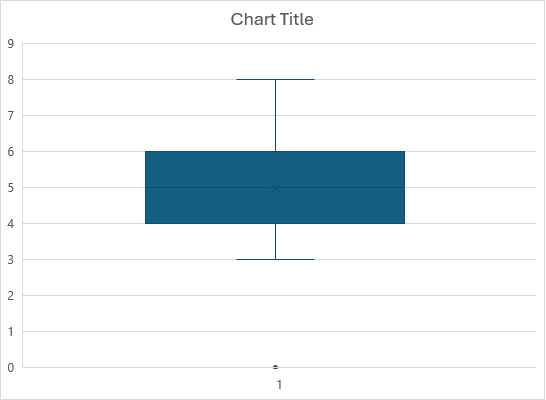
1. Histogram

From the histogram, it can be seen that the shape of the graph is almot mound shape with a single peak at 4. Hence, it is unimodal. It has potential outliers. From the shape, the graph seems to be slightly skewered to the right.

1. PIE Chart
2. Line Chart
3. Dotplot
4. Stem and leaf plot

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  | Leaf unit=1 | | | | | |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  | Number of tests taken | | | | | | | | | |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| **0** | 0 | 3 | 3 | 3 | 3 | 3 | 4 | 4 | 4 | 4 | 4 | 4 | 4 |  |  |  |  |
|  | 5 | 5 | 5 | 5 | 5 | 6 | 6 | 6 | 6 | 6 | 6 | 7 | 7 | 7 | 8 | 8 | 8 |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  | 0 | 3 | =3 | |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |

6. Box plot



From the box plot as well, we find the following:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Minimum value | First Quartile  Q1 | Median | Third Quartile  Q3 | Maximum value |
| 0 | 4 | 5 | 6 | 8 |

It has an outlier at 0, shown by an asterisks.

Calculation Chart

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Serial no of student | Number of mock test taken (x) | (x¡-x̄) | (x¡-x̄) Rounded | (x¡-x̄)^ 2 | (x¡-x̄)^ 2 Rounded | Z-score | Z-score rounded |
| 1 | 5 | 0.033333333 | 0.03 | 0.001111111 | 0 | 0.016484 | 0.02 |
| 2 | 3 | -1.966666667 | -1.97 | 3.867777778 | 3.87 | -1.08242 | -1.08 |
| 3 | 8 | 3.033333333 | 3.03 | 9.201111111 | 9.2 | 1.664835 | 1.66 |
| 4 | 4 | -0.966666667 | -0.97 | 0.934444444 | 0.93 | -0.53297 | -0.53 |
| 5 | 6 | 1.033333333 | 1.03 | 1.067777778 | 1.07 | 0.565934 | 0.57 |
| 6 | 5 | 0.033333333 | 0.03 | 0.001111111 | 0 | 0.016484 | 0.02 |
| 7 | 3 | -1.966666667 | -1.97 | 3.867777778 | 3.87 | -1.08242 | -1.08 |
| 8 | 7 | 2.033333333 | 2.03 | 4.134444444 | 4.13 | 1.115385 | 1.12 |
| 9 | 4 | -0.966666667 | -0.97 | 0.934444444 | 0.93 | -0.53297 | -0.53 |
| 10 | 4 | -0.966666667 | -0.97 | 0.934444444 | 0.93 | -0.53297 | -0.53 |
| 11 | 6 | 1.033333333 | 1.03 | 1.067777778 | 1.07 | 0.565934 | 0.57 |
| 12 | 5 | 0.033333333 | 0.03 | 0.001111111 | 0 | 0.016484 | 0.02 |
| 13 | 3 | -1.966666667 | -1.97 | 3.867777778 | 3.87 | -1.08242 | -1.08 |
| 14 | 8 | 3.033333333 | 3.03 | 9.201111111 | 9.2 | 1.664835 | 1.66 |
| 15 | 4 | -0.966666667 | -0.97 | 0.934444444 | 0.93 | -0.53297 | -0.53 |
| 16 | 6 | 1.033333333 | 1.03 | 1.067777778 | 1.07 | 0.565934 | 0.57 |
| 17 | 5 | 0.033333333 | 0.03 | 0.001111111 | 0 | 0.016484 | 0.02 |
| 18 | 3 | -1.966666667 | -1.97 | 3.867777778 | 3.87 | -1.08242 | -1.08 |
| 19 | 7 | 2.033333333 | 2.03 | 4.134444444 | 4.13 | 1.115385 | 1.12 |
| 20 | 4 | -0.966666667 | -0.97 | 0.934444444 | 0.93 | -0.53297 | -0.53 |
| 21 | 4 | -0.966666667 | -0.97 | 0.934444444 | 0.93 | -0.53297 | -0.53 |
| 22 | 6 | 1.033333333 | 1.03 | 1.067777778 | 1.07 | 0.565934 | 0.57 |
| 23 | 5 | 0.033333333 | 0.03 | 0.001111111 | 0 | 0.016484 | 0.02 |
| 24 | 6 | 1.033333333 | 1.03 | 1.067777778 | 1.07 | 0.565934 | 0.57 |
| 25 | 8 | 3.033333333 | 3.03 | 9.201111111 | 9.2 | 1.664835 | 1.66 |
| 26 | 7 | 2.033333333 | 2.03 | 4.134444444 | 4.13 | 1.115385 | 1.12 |
| 27 | 3 | -1.966666667 | -1.97 | 3.867777778 | 3.87 | -1.08242 | -1.08 |
| 28 | 0 | -4.966666667 | -4.97 | 24.66777778 | 24.67 | -2.73077 | -2.73 |
| 29 | 4 | -0.966666667 | -0.97 | 0.934444444 | 0.93 | -0.53297 | -0.53 |
| 30 | 6 | 1.033333333 | 1.03 | 1.067777778 | 1.07 | 0.565934 | 0.57 |
|  |  |  |  |  |  |  |  |
|  | 149 | -3.55271E-15 | -0.1 |  | 96.94 |  |  |

1. Mean= (sum of all measurements)/ (Total number of observations)

Here, summation of all measurements= 149 and total number of observations is 30

=149/30

=4.97

1. Standard deviation

Variance, s^2= ∑(x¡-x̄)^2/ (n-1)

= 96.94/29

=3.34

Standard deviation, s= √s^2

=√3.34

=1.82

1. Range approximation of s= If range, R is about four times of standard deviation then standard deviation can be approximated as s ≈ R/4

≈(8-0)/4

≈2

Which is very similar to the calculated value of s, 1.82.

1. Arranging all the values from lowest to highest we obtain:

0 3 3 3 3 3 4 4 4 4 4 4 4 5 5 5 5 5 6 6 6 6 6 6 7 7 7 8 8 8

First quartile, Q1 position= 0.25(n+1)

=0.25(30+1)

7.75

So, first Quartile, Q1= 4+.75(4-4)

= 4

1. 3d quartile position= 0.75(n+1)

=23.25

=6+.25(6-6)

So, third quartile, Q3 =6

1. Median = Taking the 15th and 16th value from the data series arranged from smallest to largest, we obtain the median as (5+5)/2= 5
2. Mode= From the frequency histogram, we obtain that 4 mock tests have been taken the maximum amount of times (7), hence the mode is 4
3. 70th percentile = 0,3,3,3,3,3,4,4,4,4,4,4,4,5,5,5,5,5,6,6,6,6,6,6,7,7,7,8,8,8

n (number of data points) = 30

Rank of the percentile, R= (p/100)​×(*N*+1) =0.7×31=21.7

Splitting R into the integer part (k) and the fractional part (d), we obtain

*k*=⌊21.7⌋=21,*d*=21.7−21=0.7

Here, *X*21​=6, *X*22​=6

Percentile=6+0.7×(6−6)=6

So, the 70th percentile of this data set is 6.

1. Z-score- Formula for calculation of z-score is (x-x̄)/ s

Previously we found, mean ̄x= 4.97

Standard deviation, s=1.82

The individual z-scores of each of these measurements are shown in the table below.

|  |  |
| --- | --- |
| (x¡-x̄) | Z-score  (x¡-x̄)/1.82 |
| 0.03 | 0.016484 |
| -1.97 | -1.08242 |
| 3.03 | 1.664835 |
| -0.97 | -0.53297 |
| 1.03 | 0.565934 |
| 0.03 | 0.016484 |
| -1.97 | -1.08242 |
| 2.03 | 1.115385 |
| -0.97 | -0.53297 |
| -0.97 | -0.53297 |
| 1.03 | 0.565934 |
| 0.03 | 0.016484 |
| -1.97 | -1.08242 |
| 3.03 | 1.664835 |
| -0.97 | -0.53297 |
| 1.03 | 0.565934 |
| 0.03 | 0.016484 |
| -1.97 | -1.08242 |
| 2.03 | 1.115385 |
| -0.97 | -0.53297 |
| -0.97 | -0.53297 |
| 1.03 | 0.565934 |
| 0.03 | 0.016484 |
| 1.03 | 0.565934 |
| 3.03 | 1.664835 |
| 2.03 | 1.115385 |
| -1.97 | -1.08242 |
| -4.97 | -2.73077 |
| -0.97 | -0.53297 |
| 1.03 | 0.565934 |

The value that is two standard deviations above the mean is

(x̄+2s)= 4.97+ 2\*1.82

=8.61

The value that is 1.5 standard deviations below the mean is

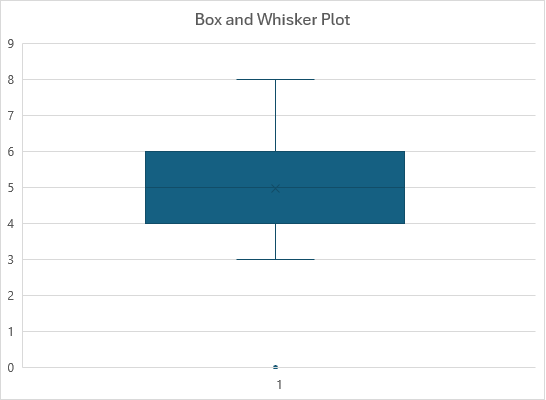
(x̄-1.5s)= 4.97- 1.5\*1.82

=2.24

From the previous calculations, the Five-Number Summary is shown in the chart below:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Minimum value | First Quartile  Q1 | Median | Third Quartile  Q3 | Maximum value |
| 0 | 4 | 5 | 6 | 8 |

From these, the Box and Whisker Plot is constructed as below:



From the chart, we find the lower fence as Q1-1.5(IQR)

= 4-1.5(6-4)

= 1

Since the value 0 lies below the lower fence, it is considered as an outlier and shown in the graph with asterisk mark.

Since the interquartile range is very small (2), and from the box and whiskers plot we see that the width of the box is very narrow as well, it appears that the middle 50% of the data is concentrated together.

Histogram displaying the data and showing the mean and +/- 1, 1.5, 2 standard deviations.

fFind *P*(−.5 ≤ *z* ≤ 1.0); *P*(−1 ≤ *z* ≤ 1); *P*(−2 ≤ *z* ≤ 2); a value of *z* having area .025 to its right (*z*.025).

A table with numbers and letters

Description automatically generated

A table with numbers and letters

Description automatically generated

From the tables, we found the area to the left of *z =* −.5 as.3085 and the area to the left of z=1 as.84134

Hence For P(−.5 ≤ *z* ≤ 1.0),

By subtracting the two entries we get

=(.84134-.30854)

=0.5328

For P(−1 ≤ *z* ≤ 1)

From the tables, we found the area to the left of *z =* −1 as.15866 and the area to the left of z=1 as.84134

By subtracting the two entries we obtain

P(.84134-.15866)

=0.68268

For P(−2 ≤ *z* ≤ 2)

From the tables, we found the area to the left of *z =* −2 as.02275 and the area to the left of z=2 as.97725

By subtracting the two entries we obtain

P(.97725-.02275)

=0.9545

For a value of *z* having area .025 to its right (*z*.025)

P(1-0.025)

=1.96

A common notation for a value of *z* having area .025 to its right is *z(*.025) = 1.96.