Inspiring Innovation and Leadership

BHR 210: Statistics Marking Scheme, June 2021 Examinations

John Karuitha

7/10/2021

Table 1: Education Levels of Employees at Karatina University

education_level	num
No_college_degree	200
Diploma	50
Bachelors_degree	200
Masters_degree	300
PhD	150

Table 2: relative Education Levels of Employees at Karatina University

education_level	proportion
No_college_degree	0.2222222
Diploma	0.0555556
Bachelors_degree	0.2222222
Masters_degree	0.3333333
PhD	0.1666667

* Also accept percentages like
22.22%, 5.56% and so on and
also fractions like 200/900,
50/900, or 2/9, 5/90 and the like

QUESTION ONE

PART A

As part of the human resource department of Karatina University, you are tasked to summarize the maximum educational level attained by each of the 900 employees at the institution. From the records, you find that 200 have no college degree (None), 50- have a diploma (DIP), 200 have a bachelors degree (BA), 300 have a masters degree (MA), and 150 have a PhD.

a) Make a frequency table with two columns; Education Level and Number. (2 marks)

The frequency table should be as in Table 1 above (2 marks)

b) Make a relative frequency table. (2 Marks)

The relative frequency should be as in Table 2 above; can either be fractions, decimals or percentages. Accept any of these formats for representing proportions.

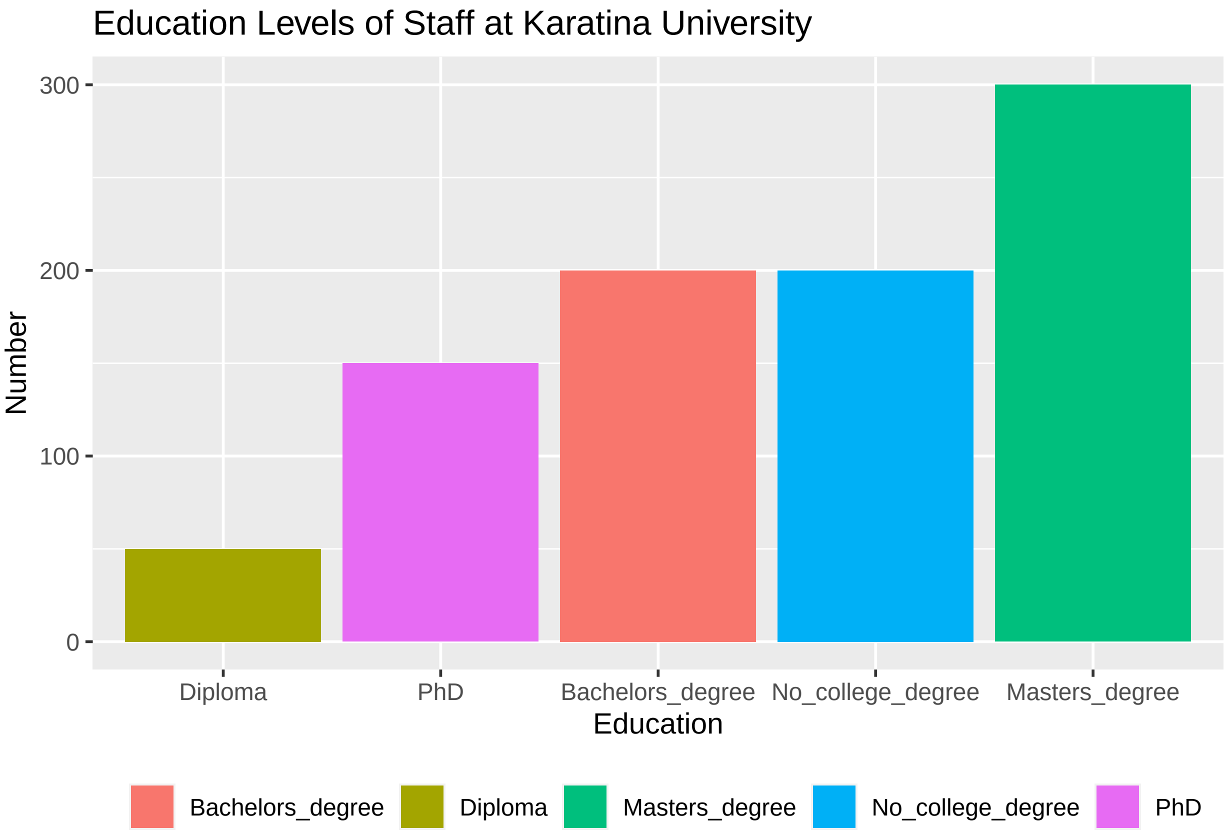
c) Is Education Level a quantitative or qualitative variable? Explain. (1 mark)

Education is qualitative or categorical as it puts the people into bins or categories and has no units.

d) Is Number a quantitative or qualitative variable? Explain. (1 marks)

Number is quantitative and discrete as it has units but cannot take on fractions: NB: The student need NOT say it is discrete, quantitative is enough for full marks.

e) Draw a bar-graph to represent the above information. (2 marks).



NB: The x-xis should represent education level with number on the y-axis.

Marks distribution: Getting the chart x and y axis right 1/2 mks Getting the bars right 1 mk. NB: The bars need NOT be ordered in ascending or descending order. The graph need NOT be colored. Titles and axis labels = 1/2 marks

f) What is the probability that an employee of Karatina University picked at random would be a holder of a diploma qualification? (1 mark)

$\frac{50}{900} = \frac{5}{90} = 0.0556$

PART B

The table below shows the corresponding height(in cm) and weight (in kilograms) of a sample of 10 traders at Karatina open air market.

Name	Wairimu	Oloo	Kiptoo	Ann	Etyang	Paul1	Jane	Joan	Paul	Carol
Height	165	172	167	179	150	140	180	200	176	220
Weight	67	73	69	75	64	50	78	84	80	90

a) Compute the arithmetic mean for the height and weight of the traders. (1 mark)

The arithmetic mean is the sum of all the respective values divide by count.

mean of height is (165 + 172 + 167 + 179 + 150 + 140 + 180 + 200 + 176 + 220) / 10 = 174.9cm.

mean of weight is (67 + 73 + 69 + 75 + 64 + 50 + 78 + 84 + 80 + 90) / 10 = 73 kgs.

b) Compute the median height and weight of the traders. (1 mark)

For the median, we first arrange the data in ascending or descending order and get the value in the middle.

Arranging height in ascending order, we get: 140, 150, 165, 167, 172, 176, 179, 180, 200, 220. The two middle values are 172 and 176. The mean of these gives us the median. (172+176)/10 = 174. (1/2 marks).

Arranging height in ascending order, we get: 140, 150, 165, 167, 172, 176, 179, 180, 200, 220. The two middle values are 172 and 176. The mean of these gives us the median. (172+176)/10 = 174. (1/2 marks).

Arranging height in ascending order, we get: 50, 64, 67, 69, 73, 75, 78, 80, 84, 90. The two middle values are 73 and 75. The mean of these gives us the median. (73+75)/10 = 74.' (1/2 marks).

c) When is the median an appropriate measure of center than the arithmetic mean? Explain. (1 mark)

The median is appropriate when the data has extreme values and/or is skewed.

d) Draw a scatter plot of the height (x - axis) and weight (y -axis). (4 marks)

- *Chart title and labels (1/2 mark each). - X and Y axis (1 mark). - Points in the chart (1.5 marks).*

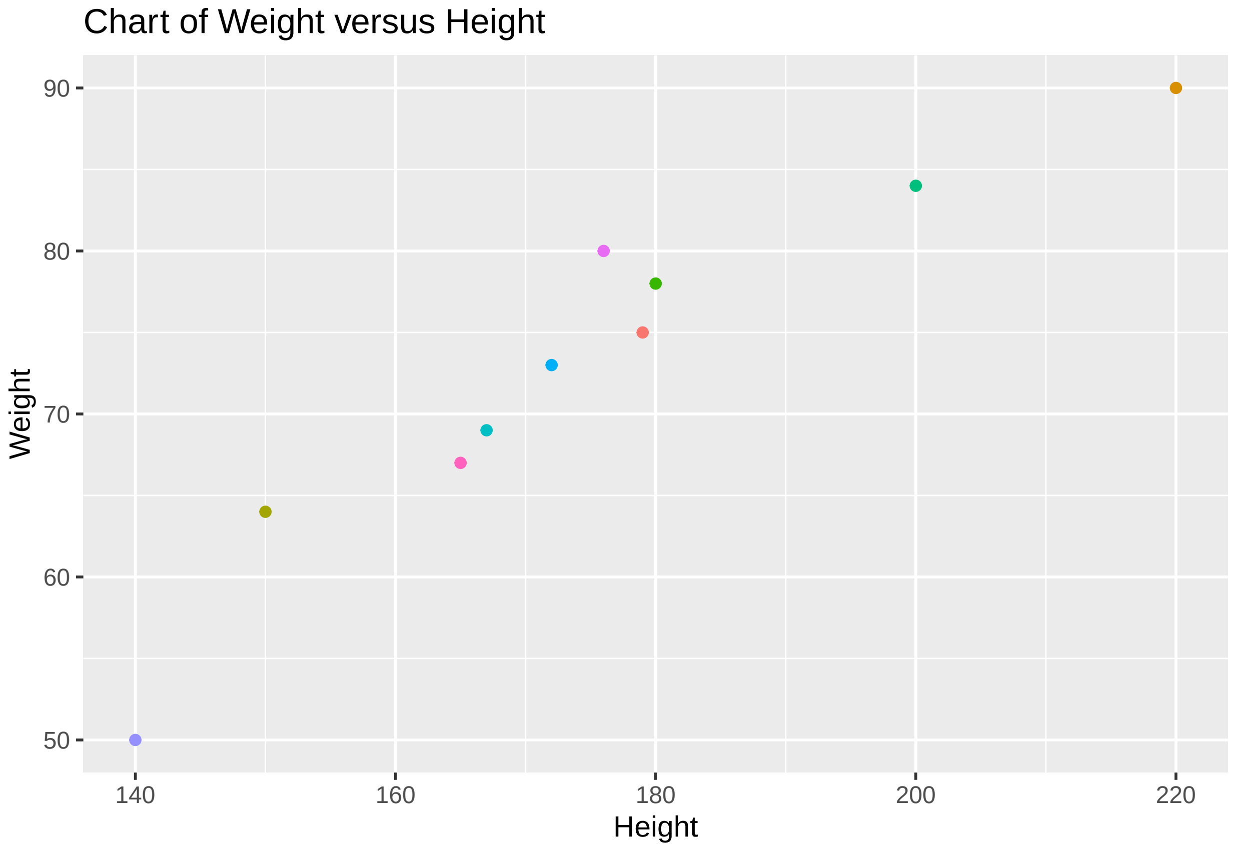


Figure 1: Scatter Plot of Height versus Height

e) Compute the standard deviation for (i) height and (ii) weight of the traders. (2 marks)

$s = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2}$

We have already computed the mean of height as 174.9 inches and weight as 73 kg above. let us denote mean of height as W and the mean of height as H we can do the table below‘

names	height	weight	height_less_H	height_less_H_sq	weight_less_W	weight_less_W_sq
Wairimu	165	67	-9.9	98.01	-6	36
Oloo	172	73	-2.9	8.41	0	0
Kiptoo	167	69	-7.9	62.41	-4	16
Ann	179	75	4.1	16.81	2	4
Etyang	150	64	-24.9	620.01	-9	81
Pauli	140	50	-34.9	1218.01	-23	529
Jane	180	78	5.1	26.01	5	25
Joan	200	84	25.1	630.01	11	121
Paulo	176	80	1.1	1.21	7	49
Carol	220	90	45.1	2034.01	17	289

*

For height, we get the sum of height less the mean of height squared (height_less_H_sq), then we divide by N-1. (table- 1 mark, solution, 1 mark)

```
## [1] 22.88838
```

Likewise, for height, we get the sum of weight less the mean of weight squared (weight_less_H_sq), then we divide by N-1. (table 1 mark, solution 1 mark)

```
## [1] 11.30388
```

f) Compute the correlation coefficient between height and weight of the traders. (3 marks)

We have already computed the standard deviations of height and weight. We also know the respective means. Hence we standardize both height and weight. using the formula.

$z_x = \frac{x-\bar{x}}{\sigma_x}$

```
##      names height weight    z_height   z_weight
## 1  Wairimu    165     67 -0.43253395 -0.5307910
## 2     Oloo    172     73 -0.12670187  0.0000000
## 3   Kiptoo    167     69 -0.34515336 -0.3538607
## 4      Ann    179     75  0.17913022  0.1769303
## 5   Etyang    150     64 -1.08788843 -0.7961866
## 6    Pauli    140     50 -1.52479142 -2.0346990
## 7     Jane    180     78  0.22282052  0.4423259
## 8     Joan    200     84  1.09662649  0.9731169
## 9    Paulo    176     80  0.04805933  0.6192562
## 10   Carol    220     90  1.97043246  1.5039080
```

We then multiply z_height with z_weight, as follows. We divide the sum of z_height_weight with standard deviation of X * standard deviation of Y.

$corr = \frac{\sum(Z_x Z_y)}{N-1}$

```
##      names height weight    z_height   z_weight z_height_weight
## 1  Wairimu    165     67 -0.43253395 -0.5307910      0.22958515
## 2     Oloo    172     73 -0.12670187  0.0000000      0.00000000
## 3   Kiptoo    167     69 -0.34515336 -0.3538607      0.12213621
## 4      Ann    179     75  0.17913022  0.1769303      0.03169357
## 5   Etyang    150     64 -1.08788843 -0.7961866      0.86616215
## 6    Pauli    140     50 -1.52479142 -2.0346990      3.10249156
## 7     Jane    180     78  0.22282052  0.4423259      0.09855928
## 8     Joan    200     84  1.09662649  0.9731169      1.06714578
## 9    Paulo    176     80  0.04805933  0.6192562      0.02976104
## 10   Carol    220     90  1.97043246  1.5039080      2.96334905
```

```
## [1] 0.9456538
```

- table 1.5 mark, solution 1.5 mark. NB: Students may use different formulaes.

g) Write a sample R code that you would use to capture the height of the traders. (2 marks)

```
height <- c(165, 172, 167, 179, 150, 140, 180, 200, 176, 220)
```

```
NB: Some students may use = instead of <- and they are right. They may also use a different variable name
```

h) Write a sample R code that you would use to compute the mean of the height (1 mark)

```
mean(height) or mean(c(165, 172, 167, 179, 150, 140, 180, 200, 176, 220))
```

QUESTION TWO

A tire manufacturer believes that the tread life of its snow tires can be described by a Normal model with a mean of 32,000 miles and a standard deviation of 2500 miles.

a) Compute the z-score for a tire that lasts for 40,000 miles. (3 marks)

$z_x = \frac{x-\bar{x}}{\sigma_x}$ (1 mark)

z = (40000-32000) / 2500 = 8000/2500 = 3.2 (2 marks)

b) Approximately what fraction of these tires can be expected to last less than 30,000 miles? (6 marks)

z = (30000-32000) / 2500 = -2000/2500 = -0.8 (3 marks)

The area less than -0.8 is equal to area over 0.8.

From our table, the area between 0 and 0.8 is equal to the area between 0 and -0.8. That area is 0.2881

Hence, 0.5-0.2881 = 0.2119

(3 marks)

c) Approximately what fraction of these tires can be expected to last between 30,000 and 35,000 miles? (6 marks)

z = (30000-32000) / 2500 = -2000/2500 = -0.8 (2 marks)
z = (35000-32000) / 3000 = 3000/2500 = 1.2 (2 marks)

As shown earlier the area between zero and -0.8 is 0.2881. The area between zero and 1.2 is 0.3849. We add the two to get 0.673. (2 marks).

QUESTION THREE

A recent study of Kenya Revenue Authority (KRA) audits showed that, for estates worth less than 15 million, about 1 out of 7 of all estate tax returns are audited, but that probability increases to 0.5 for estates worth over Ksh. 15 million. Suppose a tax accountant has three clients who have recently filed returns for estates worth more than Ksh. 15 million. What are the probabilities that:

a) All three will be audited? (3 marks)

We use the multiplication rule assuming independence.

0.5*0.5*0.5 = 0.125

b) None will be audited? (3 marks)

We use the complement rule

1 - 0.125 = 0.875

c) At least one will be audited? (5 marks)

Here, we examine 3 probabilities

P(1 is audited) + p(2 are audited) + p(3 are audited) (1 Mark)

0.5 + (0.5*0.5) + (0.5*0.5*0.5) = 0.875 (4 Marks)

d) What did you assume in calculating these probabilities? (4 marks)

- Independence

QUESTION FOUR

Is there a relationship between total team salary and the performance of teams in the Kenya National Soccer League (NFL)? For the 2019–2020 season, a linear model predicting Wins (out of 16 regular season games) from the total team Salary (Ksh. millions) for the 20 teams in the league is:

Wins = -16.32 + 0.219Salary

a) What is the explanatory or independent variable? (1 mark)

`Salary`

b) What is the response or dependent variable? (1 mark)

`Wins`

c) What does the slope mean in this context? (2 marks)

for a unit increase in salary, Wins/perfromance increase by 0.219 units (2 marks).

d) What does the y-intercept mean in this context? Is it meaningful? Explain. (4 marks)

At zero salary, the performance would be -16.32. In other words the team would loose all matches as we have 16 of them. It is meaningful as no team would exist without spending. (4 marks).

e) If one team spends Ksh. 10 million more than another on salary, how many more games on average would you predict them to win? (3 marks)

Assume a team A spends zero, winning -16.32 matches or losing 16.32 matches, and team B spends 10 million (in this case we use 10 as figures are in millions), wining -16.32 + 0.219*10 = -16.32 + 2.19 = -14.13 matches (or losing 14.23 matches). The difference is -14.23–16.32 = 2.09. Approximately 2 matches. (3 marks).

f) If a team spent ksh. 120 million on salaries and won 8 games, would they have done better or worse than predicted? (4 marks).

At 120 million (120), they should win -16.32+0.219*120 = 9.96. (3 marks) Having won 8 games, they are under performing. (1 mark).

QUESTION FIVE

a) Distinguish between a sample and population. (3 marks)

Sample: Subset of a population. Population: Large collection of individuals or objects that are the main focus of scientific enquiry.

NB: language may vary:

b) Compare and contrast the arithmetic mean and the geometric mean, giving situations where each is a more appropriate measure of center. (4 marks).

-The computation is different. Mean is the sum of the quantities of interest divided by the count of these subjects. (A fromula definition is also fine).

- The geometric mean is the nth root of the product of the quantities of interest, where n is the count of these quantities.

$Arithmetic_{m}ean = \frac{\sum x}{N}$

$geometric_{m}ean = \sqrt[n]{(X_1 * X_2 * ...... * X_n)}$

c) List and discuss four sampling techniques that researchers use to select a sample from the population (8 marks). (1 mark for listing and one for explaining)

```
- random sampling
- Convinience
- Snowballing
- Stratification
- Clustering
-   etc
```

\end{flushleft}