ANALYSIS OF THE VARIOUS ASPECTS OF THE GIVEN DATA USING NUMERICAL AND GRAPHICAL SUMMARIES AND GENERATING LOGICAL CONCLUSIONS FROM THE DATA

Contents

1 INTRODUCTION 3
2. Materials and Methods (Experimental) 3
2.1 Representing a variable using a suitable probability distribution 3
2.2 MODEL TESTING (GRAPHICALLY AND NUMERICALLY) 4
2.3 GENDER IDENTIFICATION 5
2.4 MODEL TESTING (GRAPHICALLY AND NUMERICALLY) 7
3 RELATIONSHIP BETWEEN ISLAND AND BILL DEPTH 8
4 CONCLUSION 8

LIST OF TABLES

TABLE 1. DATA OVERVIEW 3
TABLE 2.DETAILS OF DATASET 3
TABLE 3:PROBABILITY VS BILL DEPTH 4
TABLE 4, SHOWS THE CHI-SQUARE TEST PROCESS 5
TABLE 5 PROBABILITY VS BILL DEPTH 6
TABLE 6 SHOWS THE CHI-SQUARE TEST PROCESS 7
TABLE 7 RANGE VS PROBABILITY 8

LIST OF FIGURES

FIGURE 1 PROBABILITY VS BILL DEPTH 4
FIGURE 2 Q-Q PLOT FOR BILL DEPTH 5
FIGURE 3 PROBABILITY VS BILL DEPTH 6
FIGURE 4 SCATTER PLOT : BILL LENGTH VS FLIPPER LENGTH FOR MALE 7
FIGURE 5 Q-Q PLOT FOR UNIFORM DISTRIBUTION 7
FIGURE 6 BILL DEPTH VS PROBABILITY 8

Abbreviations

RV : Random Variable

1 INTRODUCTION

The information on the local penguin population in the Palmer Station, which is on Anvers Island in the Palmer Archipelago, will be analyzed and it includes crucial fields such species, body mass, sex, and the island where it is discovered.

species	island	bill_length_mm	bill_depth_mm	flipper_length_mm	body_mass_g	sex	year
Chinstrap	Dream	52.7	19.8	197	3725	male	2007
Adelie	Torgersen	34.6	21.1	198	4400	male	2007
Adelie	Torgersen	41.5	18.3	195	4300	male	2009
Gentoo	Biscoe	49.9	16.1	213	5400	male	2009

Table 1. Data overview same as table

Species	The species of the penguin (Adelie, Chinstrap or Gentoo)
Island	The island on which the penguin lives (Biscoe, Dream or Torgerson)
Bill_Length_Mm	The length of the penguin's bill (in millimeters)
Bill_Depth_Mm	The depth of the penguin's bill (in millimeters)
Flipper_Length_Mm	The length of the penguin's flipper (in millimeters)
Body_Mass_G	The penguin's body mass (in grams)
Sex	The sex of the penguin (male or female)
Year	The year the measurements were taken

Table 2.Details of dataset same as table

The aim of this project is to draw relevant and important conclusions (e.g., Relations) using various statistical tools and computer programming from the given set of data through graphical as well as numerical summaries. In addition, a technique we will be discussed by which we can identify the gender of the penguins using a suitable probability model as well as the relationship between the physical characteristics of the penguins with the island where it is inhibiting. Moreover, the accuracy and reliability of the models that were selected will be evaluated and explained in the coming sessions.

2. Materials and Methods (Experimental)

2.1 Representing a variable using a suitable probability distribution

PROCEDURE

First, a suitable random variable its type and appropriate probability distribution based on the graph obtained has to be selected. For this, a random variable from four of these physical parameters has to be chosen. Then for each random variable, a graph is plotted with the probability of finding penguins for that particular body feature.
Let, Y = Number of penguins in the data having a specific bill depth.

Consider the first random variable Y, From the table [Table 3], a graph is plotted with the x-axis as the range of value and y axis as the probability. However, with the small range, the graph cannot be analyzed as there is a missing probability for some intervals. So, by iteration, a suitable graph is selected as shown in the below diagrams [Figure 1]. Also, plotting other variables for this populations did not give proper graph which cannot be approximated to any three of the models. **Hence, bill depth is a better variable to represent these datasets in a distribution model.**

Please note that we are plotting the graph using intervals since the RV is continuous in nature.

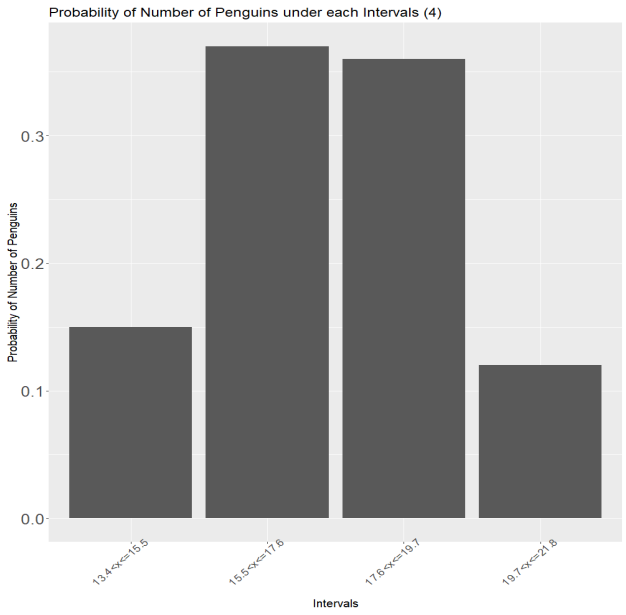


Figure 2 probability vs bill depth

range	Probability
13.4<x<=15.5	0.15
15.5<x<=17.6	0.37
17.6<x<=19.7	0.36
19.7<x<=21.8	0.12

Table 4:probability vs bill depth

From the graph, Bill depth can be approximated to normal distribution since the probability curve follows normal distribution curve. We are defining our Random variable as.

X= The number of Penguins having a given bill depth.

Our probability model is continuous RV (since the length can be measured in different accuracy and it is difficult to say the exact dimensions) and it is approximated to be Normal distribution.

ASSUMPTION

- For finding the probability distribution we categorized the data into finite interval, even though the data is a continuous one.

Mean = 17.44 cm and standard deviation =1.9 mm
Hence,
X ~ N (17.44,1.9)
N=100 ; S= 1.9 ; $\bar{x}$ = 17.44, based on 95% confidence interval we take α=0.095
So $t_{100-1,0.975} = t_{99,0.975}$=1.98
Hence

$$17.06 < \mu < 17.82 \text{ (with 95\% confidence interval)}$$

2.2 MODEL TESTING (GRAPHICALLY AND NUMERICALLY)

Null H_0 =The data set follow normal distribution.

Alternative H_1= It does not follow normal distribution.

α = 0.025 (using 95% confidence interval)

For increasing the accuracy of the test more data from the dataset is used.

lower	upper	Range	Observed Frequency (O)	Estimated frequency(E)	(O-E) ^2/E
	<13.4	<13.4	0	1.673848711	1.673849
13.4	14.4	13.4<x<=14.4	8	3.806080459	4.621279
14.4	15.4	14.4<x<=15.4	7	8.668282853	0.321075
15.4	16.4	15.4<x<=16.4	13	15.05807028	0.281288
16.4	17.4	16.4<x<=17.4	21	19.95390125	0.054843
17.4	18.4	17.4<x<=18.4	18	20.17111566	0.233688
18.4	19.4	18.4<x<=19.4	16	15.55523044	0.012717
19.4	20.4	19.4<x<=20.4	12	9.150560192	0.887302
20.4	21.4	20.4<x<=21.4	4	4.105859069	0.002729
21.4	22.4	21.4<x<=22.4	1	1.405036003	0.116762
	>22.4	>22.4	0	0.452015081	0.452015
		Total	100	100	8.657546

Table 5, shows the chi-square test process

χ^2= 8.657546

Using chi-squared test statistics, degrees of freedom=11-1-3=7,

P- value= 0.278186579 (Using R function with inputs χ^2 and degrees of freedom)

Critical value= 14.06714045

Since calculated χ^2 is less than the critical value, we fail to reject the Null Hypothesis H_O and the sample data support the normal distribution.

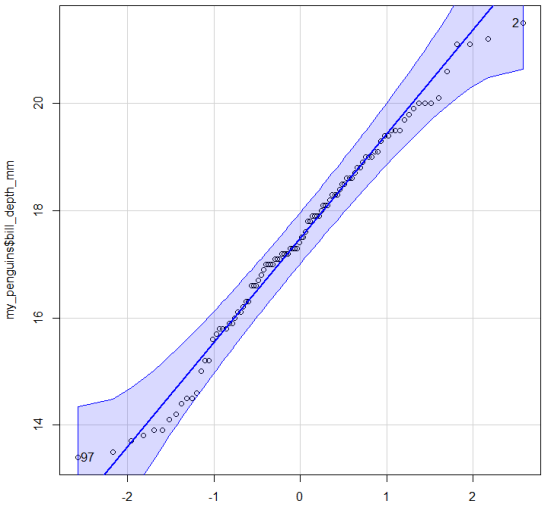


Figure 4 Q-Q plot for bill depth

The Q-Q plot as well as the numerical test suggests that the data set can be approximated to normal distribution. However, we could not get 100 % fit (this can be seen from the Q-Q plot [Figure 3]). So, a new model-variable combination may solve the problem.

Comment: **The Q-Q plot, probability graph and analytical testing show that the accuracy of the model and variable (bill depth) in identifying penguin is relatively satisfactory**.

2.3 GENDER IDENTIFICATION

For gender identification, we are following the same steps as the previous methods. Additionally, the data to be considered has been filtered to a range between 15.4 and 21.4. Using different computer iteration, **it is found that bill depth is the best factor in determining the gender of a penguin**. Here, we considered RV as the number of male penguins having a particular bill depth. The bar chart obtained is shown below [Figure 5] and based on the shape, a suitable distribution model is selected.

Bill Depth as RV

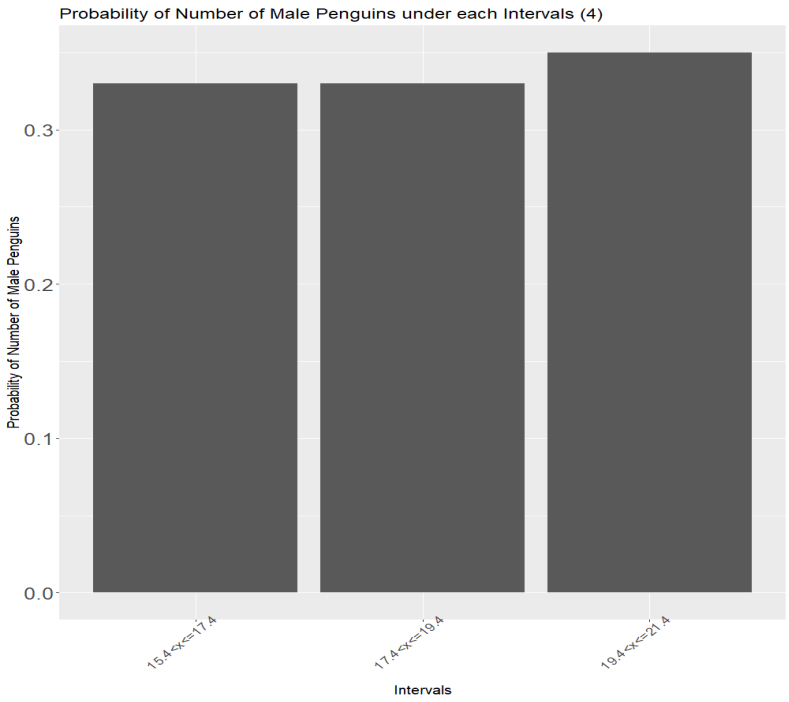


Figure 6 probability vs bill depth

Range	Probability
15.4<x<=17.4	0.33
15.4<x<=17.5	0.33
15.4<x<=17.6	0.35
Sum	1 (Rounded)

Table 6 probability vs bill depth

This graph showing the RV as Bill depth can be approximated to uniform distribution since the probability is almost same between 15.4 and 21.4. We are defining our Random variable as.

X= The number of male Penguins having a given bill depth.

The probability model is continuous RV (since the length can be measured with different accuracy and it is difficult to say the exact dimensions) and it is approximated to be a uniform distribution.

ADDITIONAL ASSUMPTION

- The probability distribution at the two ends were ignored for fitting the distribution into uniform.

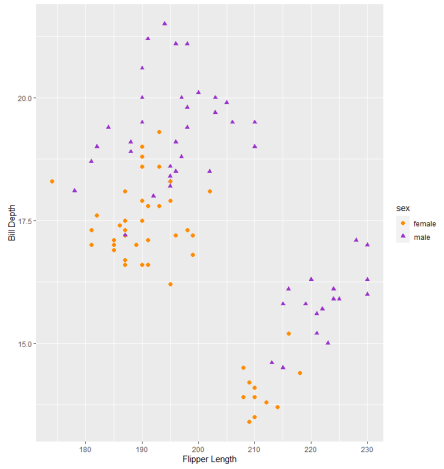
Now, we have X ~ U (17.4, 21.4). So, for our Model, the distribution is:

$$f(x) = \begin{cases} \frac{1}{6}, & 15.4 \le x \le 21.4 \\ 0, & x < 15.4 \end{cases}$$

$$\text{F (x)} = \Pr(\text{ X } \le \text{x}) = \begin{cases} 0 \\ (x - 15.4)/6 \\ 1 \end{cases} \quad \text{for x>15.4, 15.4<=x<=21.4, x>21.4}$$

E[X] = (a + b) /2 = (15.4+21.4)/2 = 18.4

Var[X] = $\frac{(b-a)^2}{12}$ = $\frac{(21.4-15.4)^2}{12}$ = 6*6/12 = 3



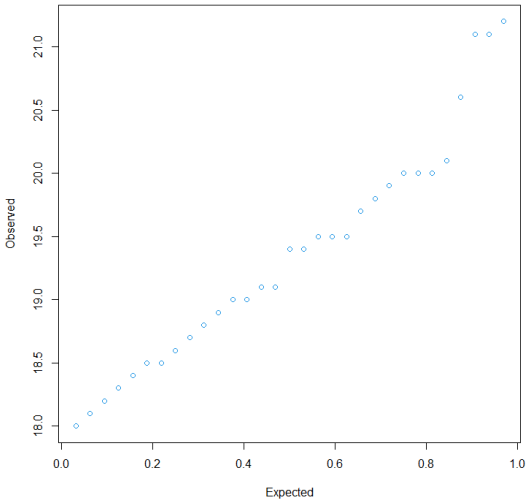
This means that **there exists around a 17 % chance that the selected penguin is male if its bill depth is between 15.4 mm and 21.4 mm. In other words, there is an 83 % chance that the selected penguin is female if it is under the above-mentioned interval. From the data set the most expected bill depth is 18.4 if it is male.**

Also, from the scatter plot [Figure 7**], we can say that penguins' having bill depth greater than 20 are likely to be males for this dataset.** Using R the correlation value between body mass and flipper length is found to be **0.86** so the importance of body mass is not that significant.

Figure 8 Scatter Plot : Bill Length vs flipper length for male

2.4 MODEL TESTING (GRAPHICALLY AND NUMERICALLY)

Quantile-Quantile plot is drawn against the RV for uniform distribution, and the following results can be seen.



The graph [Figure 9] shows the data of male penguins' bill depth between 15.4 and 21.4 and is plotted in a quantile-quantile graph to show whether it fits the selected model. Even though we could not obtain perfect matching, a uniform distribution model can approximate the given distribution.

Figure 10 Q-Q plot for uniform distribution

HYPOTHESIS TESTING OF UNIFORM MODEL.

Null H_0 = All probabilities are equal.

Alternative H_1= At least one probability is different.

$\alpha = 0.025$ (using 95% confidence interval)

Range	Observed Frequency (O)	Estimated frequency(E)	(O-E) ^2/E
15.4<x<=17.4	15	15.33333333	0.007246
15.4<x<=17.5	15	15.33333333	0.007246
15.4<x<=17.6	16	15.33333333	0.028986
Total	46	46	0.043478

Table 7 shows the chi-square test process

$\chi^2 = 0.043478$

Using chi-squared test statistics, degrees of freedom=n-1=3-1= 2,

P- value= 0.97849546 (Using R function with inputs χ^2 and degrees of freedom)
Since p> α,
we fail to reject the Null Hypothesis H_O and the sample data support the uniform distribution.
Comments: **The Q-Q plot, probability graph and analytical testing show that the accuracy of the model and variable (bill depth) in determining the gender is relatively satisfactory**.

3 RELATIONSHIP BETWEEN ISLAND AND FLIPPER LENGTH

For finding a relationship between the island and the physical properties, data has been filtered based on Adeline species because this is the only one which is common in all these islands. The suitable choice of variable can be found by plotting the histogram of four different physical parameter with same species (Adeline) and with the same island. From the probability graph, **it appears that the Flipper length is useful in determining the island where the penguin is found** since it showed a right-skewed distribution which can be approximated to **normal distribution**. (This method is followed since the plots such as bill length-flapper length, and bill depth-flapper length have not provided an appropriate graph).

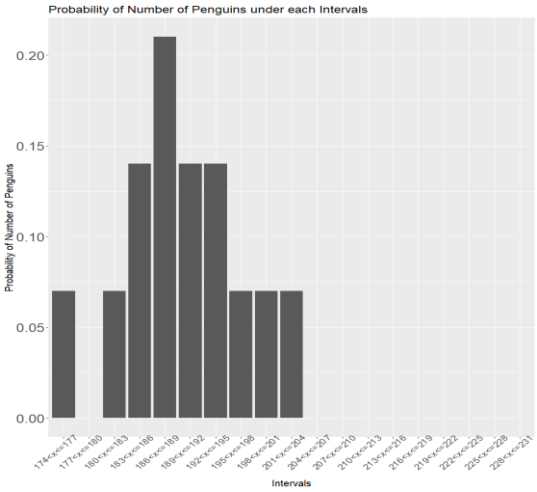


Figure 11 Flipper length vs probability

Range	Probability
174<x<=177	0.07
177<x<=180	0
180<x<=183	0.07
183<x<=186	0.14
186<x<=189	0.21
189<x<=192	0.14
192<x<=195	0.14
195<x<=198	0.07
198<x<=201	0.07
201<x<=204	0.07
204<x<=231	0

Table 8 Range vs probability

4 CONCLUSION

The data of 100 penguins is used for creating a probability model which could conditionally predict the various results based on Bill depth of the penguin. **The normal distribution** probability model is assigned to the data set with the random variables as **bill depth** related to the general population as well as for the populations of the male penguins. The range **17.06 < μ < 17.82** is estimated as a population mean for the general population data. The hypothesis test showed that we cannot reject our probability model assumption. In the second part, we tried to find a relation between bill depth and gender and for that **uniform distribution model** is selected and confirmed it using the hypothesis Test. The result can be stated as "**If the chosen penguin's bill depth is between 15.4 mm and 21.4 mm, there is a roughly 17% chance that it is a male. In other words, if the picked penguin falls within the previously given range, there is an 83% likelihood that it is female**". According to the data, a male's bill is expected to have a depth of **18.4 mm**. The standard deviation for this filtered dataset is **1.73**. **Flipper Length can be used for identifying island-physical feature relationships.**