Team Control Number

23090436

Problem Chosen

E

2023

IMMC

Summary Sheet

The article aims at developing criteria that enable one to predict biological species and sex of lizards. Data containing measurements and sexes of 564 lizards of 8 species are given. The Bayes method and Pearson Chi-Square Test are employed to indicate the criterion's effectiveness.

Firstly, we separate Species #5 from the rest. When distinguishing the biological species, we initially only use one variable, FPNr, and get a correct rate of 99.91% and a χ^2 value of 540.437. Then we decide to employ one additional variable, MBS, to classify Species # 5, since classification based on a single variable is insufficiently reliable. To emphasize the distinctions between Species #5 and other species, we develop a function to represent each individual lizard. Finally, we get a χ^2 value of 564.00 and a correct rate of 100.00%.

Next, we set up a criterion based on the ratios of some linear sizes to distinguish male from female lizards. After analyzing all the data, we discover that most variable ratios follow the normal distribution. Hence, we select 5 variable ratios whose R-square is relatively high. Then, we add weights to each ratio to emphasize their importance. Finally, we get a χ^2 value of 242.668 and a correct rate of 90.75%.

Lizards tend to live in subgroups in the same area, so it is practical to distinguish lizard species in the same area. We classify Species #6 and #7 in a subgroup and get a χ^2 value of 143.00 and a correct rate of 100.00%; Species #1 and #2 in a subgroup and get a χ^2 value of 92.08 and a correct rate of 92.24%. We then classify Species #3, #4 and #5 in a subgroup. Since we have classified Species #5 with 100.00% accuracy, we only need to distinguish Species #3 from #4 in this subgroup. And we obtain the χ^2 value of 209.14 and a correct rate of 97.57%.

Finally, we develop a set of criteria to classify the 564 lizards in terms of species and sex. The order and the criterion to classify them must be reasonably established. In order to reduce the subsequent classification mistakes brought by the original classification errors, we need to make early classification processes sufficiently accurate. We arrange the classification order according to their classification accuracy. The result classification order is: Species #5, #4, #7, the subgroup of species #1 and #2, Species #3, and then Species #6. After that, all lizards left are classified as Species # 8. Therefore, we do an additional modification to the classification of Species # 8. The final correct rate is 85.83%.

Lizard Sex and Species Classification

Team#23090436

January 17, 2023

Contents

1 Introduction

1.1 Background

Evolutionary biologists typically use reproductive compatibility as a major criterion for distinguishing among different species. In other words, members of two different species usually cannot mate, or if they do, the offspring are usually sterile, unable to survive, or are less fit than normal. This reproductive incompatibility is a primary factor in species differentiation and provides a reliable way to determine species boundaries. Biologists are actively seeking more efficient methods of differentiating among species. For example, they hoped to distinguish between two species based solely on their body shapes or habitats. This could save a considerable amount of time and effort in identifying species, and help to provide a more accurate understanding of the biodiversity of a given area. Furthermore, this kind of research could help to inform conservation efforts and research into the interactions between different species. Despite the various visible characteristics that can be used to distinguish among different species, it is not always easy to differentiate between two closely related species. This is often the case when subtle differences exist, such as variations in skin color, leg length, and other traits. As a result, it can be challenging to differentiate between two closely related species based solely on the more obvious criteria mentioned above. According to a zoological journal [1], there are three typical kinds of evidence used to classify species, behavioral, ecological and molecular data. Behavioral data refer to nesting habits and foraging ecology of the species, which can be expressed in numbers for easier statistics. Ecological data include the living condition and habitat distribution of every species. And molecular data can be analyzed by DNA data of the species.

1.2 Problem Restatement

We are required to develop criteria that can predict the sex and biological species with the highest possible accuracy based on measurements of 564 lizards of 8 species. The measurements include the numbers of scales on lizards' bodies and linear sizes of the lizards' body parts, which can show the pholidosis characteristics and morphometric characteristics of lizards. Meanwhile, the criteria should be obtainable and relatively obvious. And specific requirements are shown below:

1. Set up a criterion only based on one variable (femoral pore number on the right side) to distinguish Species #5 from other lizards with the highest possible accuracy.

2. Set up a criterion based on two variables respectively from the pholidosis characteristics and morphometric characteristics to distinguish Species #5 from other lizards with a highest possible accuracy.

3. Set up a criterion to predict the sex of lizards, regardless of biological species, based on the ratios of some given measurements with the highest possible accuracy.

4. Set up a set of criteria to distinguish all species in the given 3 subgroups with the highest possible accuracy:

    - Species #6 and #7

- Species #1 and #2
- Species #3, #4 and #5

5. Set up a set of criteria to predict the sex and species of lizards in the whole group with a highest possible accuracy.

1.3 Our Work

1. **For Task 1, we:**
   (a) analyze the distribution of Species #5's FPNr;
   (b) establish a performance metric;
   (c) find an optimal criterion to classify Species #5 from others.
2. **For Task 2, we:**
   (a) set up a standard to select predicting variables;
   (b) develop a function based on the best pair of predicting variables;
   (c) find an optimal criterion to classify Species #5 from others.
3. **For Task 3, we:**
   (a) establish a function to reflect the effectiveness of the variable ratio;
   (b) analyze the normal distribution of each variable ratio;
   (c) add weights to each ratio and develop a predicting criterion.
4. **For Task 4**, we developed the criterion (criteria) to distinguish the species in each group based on the standard of finding the best predicting variables.
5. **For Task 5**, we:
   (a) determine the classification order to minimize the subsequent classification errors;
   (b) simplify the classification as much as possible to reduce the complexity of the calculation.

2 Notation

Table 1: Notation

Symbol	Definition
P_c	The correct rate of a classification result
$j_i(a,b)$	The number of Species #i lizards whose j values range from a to b
σ_{j_i}	Species #i lizards' standard deviation of j
$\overline{j_i}$	Species #i lizards' average value of j
$\sigma_{M_{p,q}}$	The male lizards' standard deviation of $\frac{p}{q}$
$\sigma_{F_{p,q}}$	The female lizards' standard deviation of $\frac{p}{q}$
$\overline{M_{p,q}}$	The male lizards' average value of $\frac{p}{q}$
$\overline{F_{p,q}}$	The female lizards' average value of $\frac{p}{q}$

3 Lizards' Species Partial Classification

3.1 Species Classification based on FPNr

This model aims at distinguishing Species #5 from other species, while referencing only one characteristic, FPNr. We primarily calculate the average value of FPNr for each species, and the results are shown below:

Table 2: Average value of FPNr for each species

Species num	Average value of FPNr
1	18.424
2	17.127
3	15.378
4	18.699
5	8.958
6	15.867
7	18.045
8	17.250

Meanwhile, we draw a box plot to illustrate the distribution of lizards of Species #5 and not of Species #5 in terms of FPNr, as shown below:

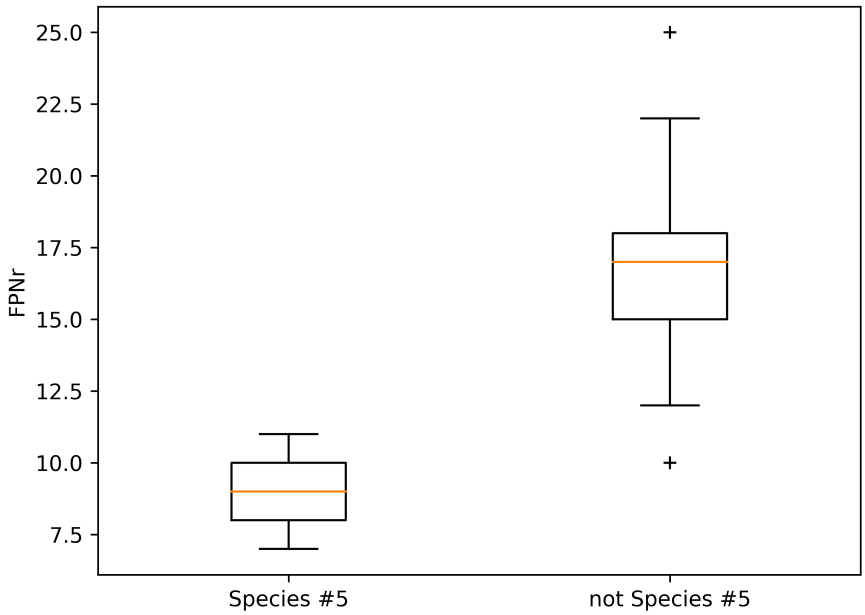

Figure 1: The box plot of FPNr distribution

We can tell from the table above that the FPNr of Species #5 differ significantly from those of non-Species #5. Hence, we can consider that a lizard can be classified as Species #5 if its FPNr value is below a certain level (and vice versa). Consequently, setting the threshold N is the following step. In other words, if a lizard's FPNr value is below N, it can be classified as Species #5.

A 2×2 Contingency Table of the classification results produced by each criterion can be created, and the chi-square value χ^2 can then be calculated.

Table 3: Classification result of a criterion

	Species #i	Not Species #i
Classified as Species #i	a	b
Classified as not Species #i	c	d

χ^2 of the result above can be calculated as:

$$\chi^2 = \frac{(a+b+c+d)(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)} \tag{3.1}$$

The χ^2 value can indicate whether or not our classification is correlated (positively or negatively) to the real distribution of species. The value of χ^2 can be considered as a reference to the correct rate: when the correct rate is relatively high, the higher the χ^2 value, the better the criterion is.

We can also calculate the correct rate of the result above. $\frac{a}{a+c}$ (the probability of a lizard classified as Species #5 when it is actually Species #5) and $\frac{d}{b+d}$ (the probability of a lizard classified as non-Species #5 when it is actually non-Species #5) can both reflect the accuracy of our classification. Therefore, we determine the performance metric to be the average of the two Bayes possiblities. Hence,

$$P_c = \frac{1}{2} \cdot (\frac{a}{a+c} + \frac{d}{b+d}) \tag{3.2}$$

The larger its value, the closer our classification results are to the actual situation. So each N's P_c can be considered a performance metric.

Each possible N's performance metric and χ^2 is shown below.

Table 4: Performance metrics and χ^2 of possible N

N	χ^2	**Performance Metrics**
8	67.861	0.125
9	229.061	0.417
10	346.721	0.625
11	467.345	0.875
12	540.437	0.999
13	479.847	0.993
14	357.031	0.976
15	153.890	0.898
16	61.138	0.741
17	27.686	0.548
18	12.938	0.356

From the table above, we can figure out 12 is the optimal value of N, since it has the greatest performance metric. In this case, the table below shows how the criterion performs:

Table 5: Performance of $N = 12$

	True Species #5	True Species #1-4, 6-8
Classified as Species #5	24	1
Classified as Species #1-4, 6-8	0	539

Among the 564 lizards, we mis-classify 1 lizard, which accounts for 0.18%.

3.2 Variable Selection

Since classification based on just one variable is not so reliable, we decided to create a criterion based on two variables. The initial step entails choosing the representative variables that can clearly distinguish

Species #5 from other lizards. To determine whether the variable j is suitable, the ratio r_j is defined:

$$r_j = \frac{j_5(\overline{j_5} - \sigma_{j_5}, \overline{j_5} + \sigma_{j_5})}{\Sigma_{i=1}^{8} j_i(\overline{j_5} - \sigma_{j_5}, \overline{j_5} + \sigma_{j_5})} \tag{3.3}$$

r_j indicates the extent to which a variable is capable of distinguishing Species #5 from others: the greater its value, the more suitable a variable is. Below are the r_j results for each variable:

Table 6: r_j of each variable

Pholidosis		Morphometric	
Variable(j)	$r_j(\times 10^{-2})$	**Variable**(j)	$r_j(\times 10^{-2})$
MBS	59.26	SVL	5.8
VSN	5.60	TRL	3.18
CSN	21.65	HL	9.52
GSN	51.43	PL	14.95
FPNr	94.74	ESD	11.67
SDLr	5.15	HW	5.84
SCSr	5.07	HH	7.63
SCGr	10.53	MO	12.86
SMr	3.70	FFL	11.59
MTr	6.13	HFL	26.23
PA	5.36		
PTMr	5.05		
aNDSr	5.41		

We can choose variables with a relatively high r_j based on the known outcome. Therefore, we can choose three variables (MBS, GSN, and FPNr) as the strongest predictors since each of their ratios (R) is larger than 50%, which is much higher than all other variables.

We can pick variables with relatively large r_j values based on the known outcome. Because the ratio R of the three strongest predictors, MBS, GSN, and FPNr, is each greater than 50%, which is much higher than all other variables, we can choose these three as our top candidates.

3.3 Species Classification based on MBS and FPNr

These three optional variables can be combined in pairs. After our attempts and calculations, we decide to choose MBS and FPNr as the predicting variable pairs, as they have the highest accuracy and χ^2 of the output results.

Even though the difference between Species #5 and other species in MBS and FPNr is the relatively most significant, we still need to develop a function that can magnify the differences. The scatter plot of MBS is shown below:

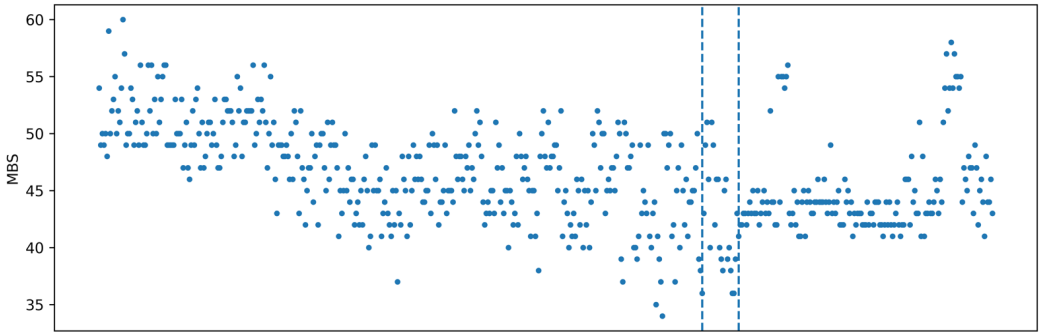

Figure 2: The scatter plot of MBS distribution

Dots in the rectangle divided by blue dashed lines in this figure refers to the MBS distribution of Species #5. We can see that MBS of Species #5 is not all that different from other species, as shown in the figure above.

Obviously, a simple linear relationship cannot adequately reflect the impact of these two variables on species classification, hence, we consider multiplying two variables. Our function can be expressed as:

$$f(x, y) = x \times y \tag{3.4}$$

In this case, x and y refer to MBS and FPNr.

The next step is to determine the threshold M: if a lizard's MBS and FPNr's function result is lower than M, it can be classified as Species #5. We go through the range of possible M values and determine that $M = 450$.

In this instance, the table below shows how the criterion performs:

Table 7: Performance of $M = 450$

	Species #5	Not Species #5
Classified as Species #5	24	0
Classified as not Species #5	0	540

χ^2 and P_c of this result can also be calculated:

$$\chi^2 = 564.000, P_c = 100.00\% \tag{3.5}$$

These two metrics are both relatively high, we can consider it as an optimal classification method.

3.4 Sex Classification

Since it is expected that the ratios of some measured linear sizes and sex are correlated, we first look at the ratios of variables as the determinants of sex.

Selecting the appropriate ratio of two variables from a pool of 23 variables is the initial stage. We create another variable to indicate the ratio, $R_{p,q}$, in order to demonstrate whether it is effective or not.

$$R_{p,q} = \frac{|\overline{M_{p,q}} - \overline{F_{p,q}}|}{\sigma_{M_{p,q}} + \sigma_{M_{p,q}}} \tag{3.6}$$

The great majority of the value $\frac{p}{q}$ will be dispersed within $(\mu - 3\sigma, \mu + 3\sigma)$ because $\frac{p}{q}$ follows a normal distribution. (μ refers to the average value and σ refers to the standard deviation) Therefore, if $R_{p,q}$ is 3, the sex classification can be considered (3σ Limits) determined by the two variables p and q. The more significant the ratio $\frac{p}{q}$ between the male and female, the greater the $R_{p,q}$ value.

After our calculation, we can get the 5 optimal $\frac{p}{q}$ of variables p and q:

Table 8: The optimal $\frac{p}{q}$ variables (linear sizes only)

Variable 1(p)	Variable 2 (q)	$R_{p,q}$
TRL	HH	1.078
TRL	HL	1.209
SVL	HL	1.296
TRL	PL	1.429
TRL	HW	1.630

We can get the frequency distribution table of each ratio. An example is shown below.

Table 9: The frequency of Male and Female in different groups of $\frac{\text{SVL}}{\text{HL}}$

Midpoint	Group range	Frequency (Male)	Frequency (Female)
1.9375	[1.833,2.042)	1	1
2.1465	[2.042,2.251)	2	0
2.3555	[2.251,2.460)	2	0
2.5645	[2.460,2.669)	0	0
2.7735	[2.669,2.878)	88	10
2.9825	[2.878,3.087)	152	66
3.1915	[3.087,3.296)	25	136
3.4005	[3.296,3.505)	4	57
3.6095	[3.505,3.714)	0	16
3.8185	[3.714,3.923)	0	1
4.0275	[3.923,4.132)	0	1
4.2365	[4.132,4.341)	0	0
4.4455	[4.341,4.550)	1	0
4.6545	[4.550,4.759)	0	0
4.8635	[4.759,4.968]	0	1

After that, we can fit the frequency distribution plot of male and female lizards. An example is shown below.

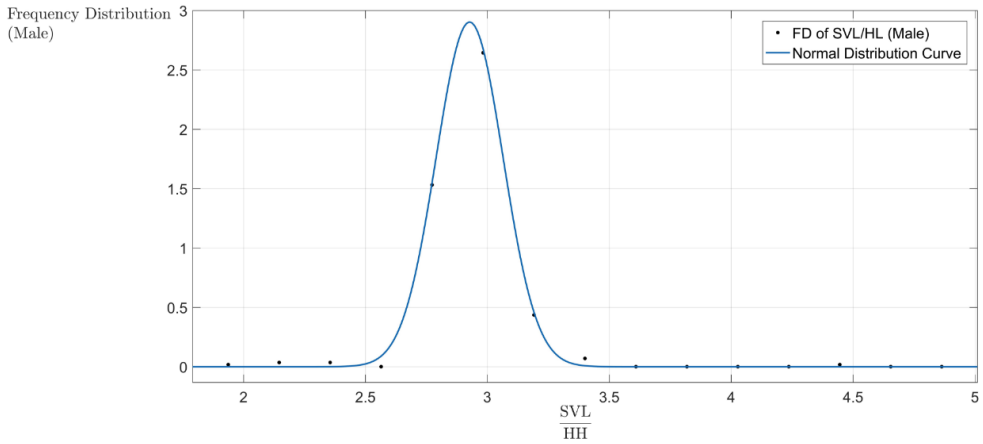

Figure 3: The normal distribution of Male's $\frac{\text{SVL}}{\text{HL}}$

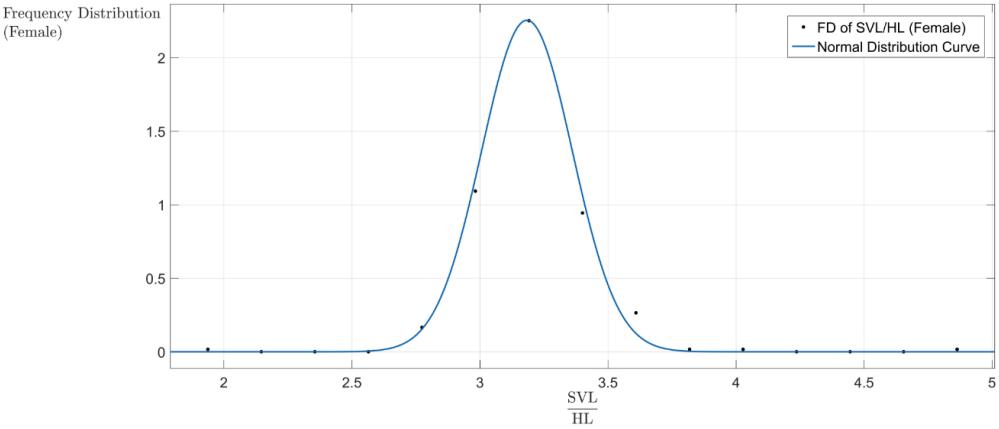

Figure 4: The normal distribution of Female's $\frac{\text{SVL}}{\text{HL}}$

We can see that the distributions of $\frac{p}{q}$ of both sexes follow normal distribution, which can be expressed as:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{3.7}$$

This analysis provides insight into the distribution of $\frac{p}{q}$ for both male and female populations. We choose the $\frac{p}{q}$ whose R-square is near 1 and since $\frac{p}{q}$ follows the normal distribution, we can use the $\frac{p}{q}$ to differentiate between sexes. Then, we identify 16 frequency distribution plots of $\frac{p}{q}$, half of which are specific

to male populations and half to female populations. The specific data of σ, μ, R-square of each variable ratio $\frac{p}{q}$:

Table 10: σ, μ, R-square of each variable ratio $\frac{p}{q}$

Variable Ratio($\frac{p}{q}$)	Sex	σ	μ	R-square
$\frac{\text{SVL}}{\text{HL}}$	M	0.1374	2.927	0.9979
	F	0.1769	3.184	0.9925
$\frac{\text{TRL}}{\text{HH}}$	M	0.6500	5.377	0.9940
	F	0.7451	6.057	0.9598
$\frac{\text{TRL}}{\text{HL}}$	M	0.1153	1.510	0.9864
	F	0.1853	1.702	0.9874
$\frac{\text{TRL}}{\text{HW}}$	M	0.3693	3.497	0.9465
	F	0.4162	4.279	0.9557
$\frac{\text{TRL}}{\text{PL}}$	M	0.1785	2.252	0.9892
	F	0.2799	2.587	0.9790

From the frequency distribution plot of the two sexes, we can calculate the probability that a lizard is male. For each lizard's $\frac{p}{q}$,

$$P_{\text{male}} = \frac{f_{\text{male}}(x)}{f_{\text{male}}(x) + f_{\text{female}}(x)}, x = \frac{p}{q}$$

We can then multiply all ratios P_{male} and set the threshold to determine the sex. The result is shown below:

Table 11: Performance of not considering weights

	Male	Female
Classified as male	271	146
Classified as female	4	143

The threshold of this result is 9.4×10^{-5}. Meanwhile, χ^2 and P_c of this result can also be calculated:

$$\chi^2 = 168.662, P_c = 74.01\%$$

Among the 564 lizards, we mis-classify 150 lizards, which accounts for 26.60%. Clearly, the best way to distinguish between male and female lizards is not to simply multiply the ratios. Therefore, we take into consideration applying weights to each variable ratio P_{male}. We let the weight of a ratio be a function of $R_{p,q}$, since ratio $R_{p,q}$ shows the significance of ratio $\frac{p}{q}$ in determining sex. We use $2 * {R_{p,q}}^{1.5}$ as the index of each ratio P_{male}.

In this context, the table below shows how the criterion performs:

Table 12: Performance of considering weights

	Male	Female
Classified as male	228	50
Classified as female	47	239

Among the 564 lizards, we mis-classify 97 lizards, which accounts for 17.20%. The threshold of this result is 8×10^{-6}. χ^2 and P_c of this result can also be calculated:

$$\chi^2 = 242.668, P_c = 82.80\%$$

This classification method is optimal, as evidenced by the fact that both of the metrics used are higher than the result of not adding weights. This suggests that the addition of weights has resulted in a more accurate and efficient set of criteria. Overall, this classification method can be considered a success.

3.5 Species Classification in Subgroups

Not all lizard species in question are found in the same locations. This means that the range of each species must be considered when conducting research or studying a particular lizard species. Consequently, in a practical setting, it is often necessary to distinguish between species and their associated subgroups that coexist. This task can be particularly challenging, as species may exhibit similar characteristics and behaviors, making it difficult to differentiate between them. Therefore, careful analyses are required in order to accurately distinguish between species. We are going to distinguish species in 3 subgroups.

3.5.1 Classifying Species #6 and #7 in a Subgroup

The data provided suggests that there is a significant difference in MBS and SCGr between Species #6 and Species #7. To magnify the differences in variables between these two species, we consider multiplying the two variables together. Hence, the function can be expressed as:

$$g(x, y) = x \times y \tag{3.8}$$

Here x and y refer to MBS and SCGr. Then we can get the following result:

Table 13: Performance of classifying Species #6 and #7

	True Species #6	True Species #7
Classified as Species #6	120	0
Classified as Species #7	0	22

The threshold of this result is 440. χ^2 and P_c of this result can also be calculated:

$$\chi^2 = 143.000, P_c = 100.00\%$$

3.5.2 Classifying Species #1 and #2 in a Subgroup

We can observe from the provided information that two species have distributions for the three variables: MTr, HFL, and PTMr, as is shown below:

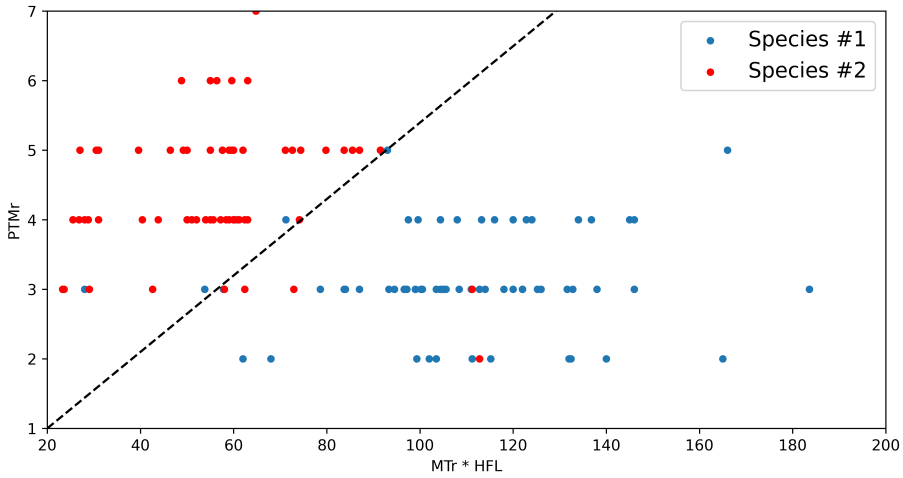

Figure 5: The distribution of Species #1 and #2

From the figure above, we can observe that Species #1 and Species #2 are relatively different. The dashed line in the figure clearly delineates Species #1 from Species #2, indicating a marked distinction between the two. Hence, we can consider our function as:

$$f(x, y, z) = \frac{xy - 20}{z - 1} \tag{3.9}$$

Here x, y, z refer to the value of MTr, HFL and PTMr. And the dashed line can be considered as the threshold: a lizard will be classified as Species #2 if the return value of the function above is less than the threshold; else, it will be classified as Species #1.

Then, we can get the following result:

Table 14: Performance of classifying Species #1 and # 2

	True Species #1	True Species #2
Classified as Species #1	61	5
Classified as Species #2	5	58

Among the 129 lizards, we mis-classify 10 lizards, which accounts for 7.75%. The threshold of this result is 18.1. χ^2 and P_c of this result can also be calculated:

$$\chi^2 = 92.08, P_c = 92.24\%$$

3.5.3 Classifying Species #3, #4 and #5 in a Subgroup

As outlined in Section 3.3, we have created an ideal criterion to differentiate Species #5 from the other species. Therefore, in this instance, our primary focus should be on how to classify Species #3 and #4. By looking at the data presented, it is evident that two species possess different distributions for the two variables of MBS and GSN, as displayed below:

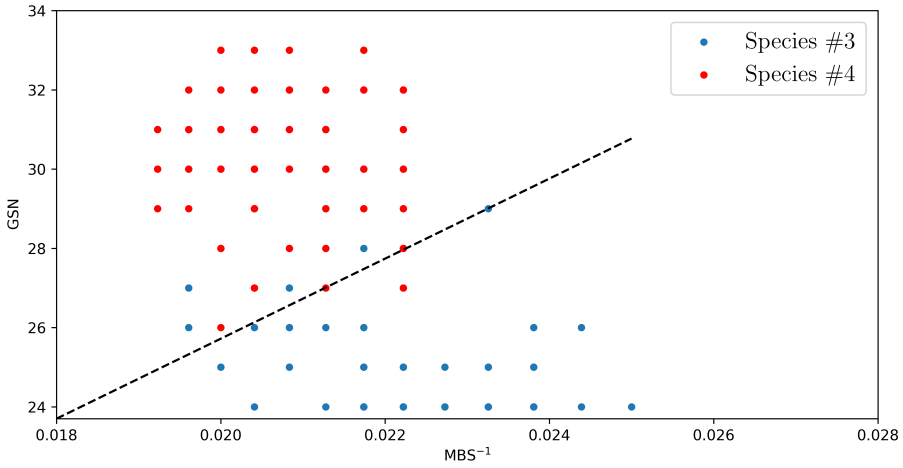

Figure 6: The distribution of Species #3 and #4

We can see from the above figure that Species #3 and Species #4 are quite diverse. The distinct dashed line in this figure, which shows a clear division between Species #3 and #4, clearly separates the two.

$$f(x, y) = \frac{x - 23.7}{y^{-1} - 0.018} \tag{3.10}$$

Here x, y refer to the value of MTr, HFL and PTMr. And the dashed line can be considered as the threshold: a lizard will be labeled as Species #3 if its result value from the aforementioned function is lower than the threshold; otherwise, it will be labeled as Species #4.

Then, we can get the following result:

Table 15: Performance of classifying Species #3, #4 and #5

	True Species #3	True Species #4	True Species #5
Classified as Species #3	148	2	0
Classified as Species #4	8	91	0
Classified as Species #5	0	0	24

Among the 273 lizards, we mis-classify 10 lizards, which accounts for 3.66%.

Table 16: Performance of classifying Species #p, #q and #r

	True Species #p	True Species #q	True Species #r
Classified as Species #p	a	b	c
Classified as Species #q	d	e	f
Classified as Species #r	g	h	i

We need to determine how to calculate the χ^2 and P_c of the 3×3 Contingency table shown above. According to the Pearson's chi-squared test, we can get that :

$$\chi^2 = \sum_{i=1}^{3}\sum_{j=1}^{3} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}} \tag{3.11}$$

And P_c can be expressed as:

$$P_c = \frac{1}{3} \cdot (\frac{a}{a+d+g} + \frac{e}{b+e+h} + \frac{i}{c+f+i}) \tag{3.12}$$

Hence, the threshold of this result is 1010. χ^2 and P_c of this result can also be calculated:

$$\chi^2 = 209.14, P_c = 97.57\%$$

4 Lizards' Species and Sex Classification

4.1 Sex Classification

For sex classification, we can follow the criterion developed in Section 3.4 to classify the sex. We also use the scales on different parts of the lizards' bodies, namely count of scales, in addition to linear sizes. We employ the same method as is used in Section 3.4, and the optimal $\frac{p}{q}$ values are listed below.

Table 17: The optimal $\frac{p}{q}$ variables (count of scales and linear sizes)

Variable 1(p)	Variable 2 (q)	$R_{p,q}$
VSN	FFL	1.168
VSN	HH	1.177
TRL	HL	1.209
SVL	HL	1.296
TRL	PL	1.429
VSN	HL	1.548
TRL	HW	1.630
VSN	PL	1.822

Employing those ratios as predicting variables, we get the following result.

Table 18: Sex classification results

	Male	Female
Classified as male	204	7
Classified as female	71	282

Among the 564 lizards, we mis-classify 78 lizards, which accounts for 13.83%. The threshold of this result is 1×10^{-6}.

$$\chi^2 = 309.894, P_c = 85.88\%$$

4.2 Species Classification

To distinguish between all of the species, we can classify lizards based on already established categorization standards. We should attempt to classify the species with high accuracy initially in order to decrease the subsequent classification error caused by the prior classification error. As a result, choosing the order in which to classify the species and simplifying the criterion are the two most important steps in this process to prevent classification errors brought on by calculating complicity.

4.2.1 To classify Species #5

We can categorize Species #5 using the standard described in Section 3.3. This step won't have an impact on the classification in the stages that follow because the correct rate for this classification is 100%. As a result, it makes perfect sense to classify it as the first of the 8 lizard species.

4.2.2 To classify Species #4

We can classify Species #4 by the value of variable GSN and PTMr. Species #4 and other species have the distribution for these 2 variables shown below:

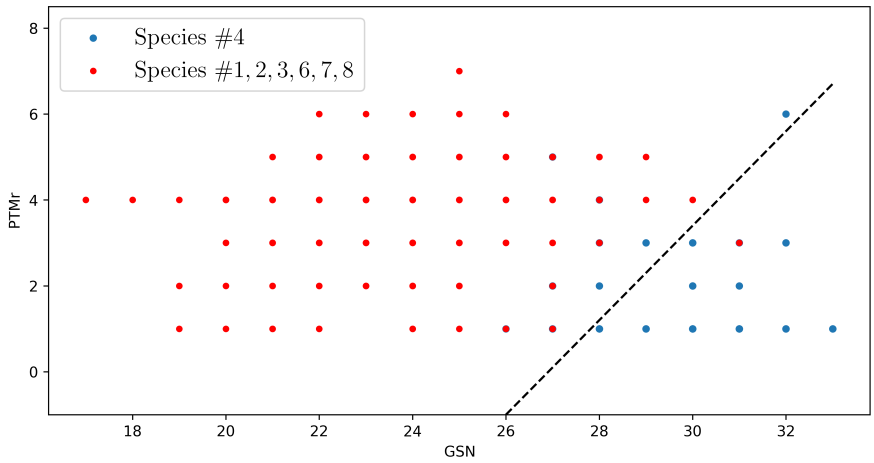

Figure 7: The distribution of Species #4 and the rest lizards

From the figure above, we can observe that Species #4 distributed quite differently from other species. Hence, we can get our formula:

$$f(x, y) = \frac{x + 1}{y - 26} \tag{4.1}$$

Here x and y refer to the value of PTMr and GSN. And the dashed line can be considered as the threshold: a lizard will be classified as Species #4 if the return value of the function above is lower than the threshold. We can establish the threshold as 1.1 after transposing the data.

4.2.3 To classify Species #7

1. Based on Sex

   According to the given data, lizards of Species #7 are all female. Therefore, we conduct a basic sex assessment of lizards which could be Species #7.

2. Based on MBS and SDLr

   After the initial screening, we observe that the ratio $\frac{\text{MBS}}{\text{SDLr}}$ differentiate from Species #7 to the rest species. Hence, we get a function shown below:

   $$f(x) = \frac{x}{y} \tag{4.2}$$

   Here x and y refer to the value of MBS and SDLr. The following step is to determine the threshold for the result: When the result value of a lizard is larger than the threshold, it can be classified as Species #7. We finally determine the threshold as 2.5.

4.2.4 To classify Species #1 and #2

1. Classify the subgroup of Species #1 and #2

   Following our analysis of the data, no distinct measurements of Species #1 or Species #2 could be identified that would set them apart from the other lizard species. Consequently, we thought it prudent to separate Species #1 and Species #2 from the rest of the species before deciding whether the lizard was a member of Species #1 or Species #2. We can classify the subgroup of Species #1 and #2 by the variable MBS and ESD. The subgroup and other species have the distribution for these 2 variables shown below:

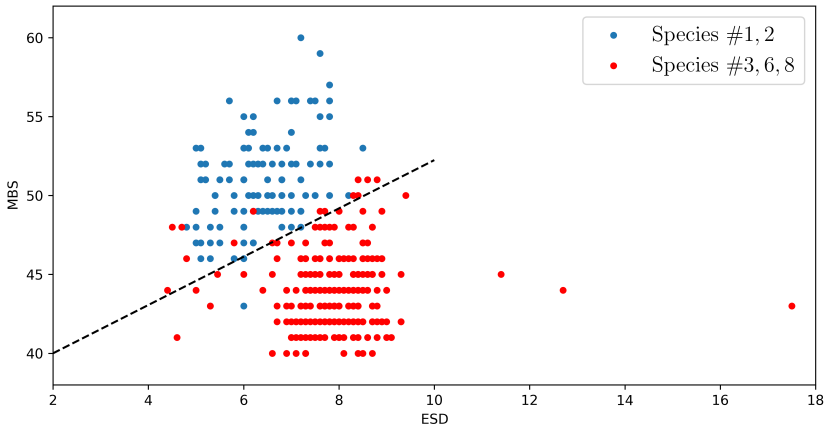

Figure 8: Classifying the subgroup of Species #1 and #2

   From the figure above, we can observe that the subgroup distributed quite different from other species. Hence, we can get our formula:

$$f(x, y) = \frac{x - 40}{y - 2} \tag{4.3}$$

   Here x and y refer to the value of MBS and ESD. Choosing the threshold for the outcome is the next step: A lizard can be categorized as the subgroup of Species #1 and Species #2 when its result value above the threshold. Finally we determine threshold as 1.53.

2. Distinguish Species #1, #2 in the Subgroup

   As outlined in Section 3.5.2, we can classify Species #1 and #2 in a subgroup with 100% accuracy.

4.2.5 To classify Species #3

Species #3 can be categorized based on the values of the variables VSN and aNDSr. The distribution for these 2 variables in Species #3 and other species is as follows:

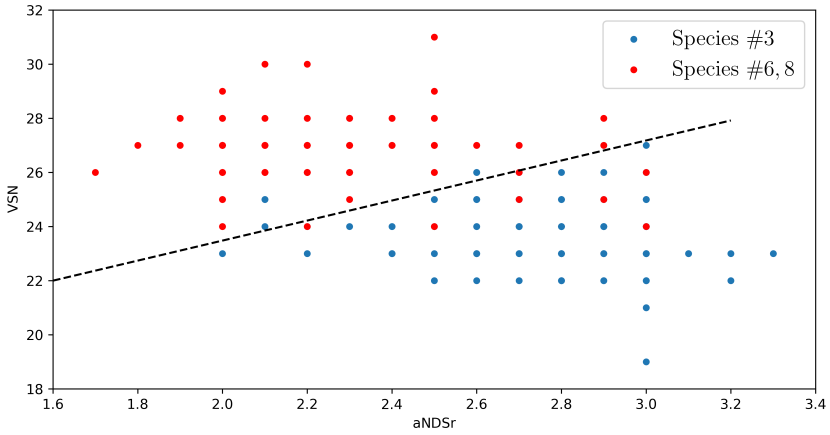

Figure 9: Classifying Species #3

We can see from the following figure that Species #3 is distributed very differently from other species. We can so arrive to our formula:

$$f(x, y) = \frac{x - 22}{y - 1.6} \tag{4.4}$$

Here, the values of VSN and aNDSr are denoted by x and y. A lizard will be categorized as belonging to Species #3 if the return value of the function above is less than the dashed line, which may be thought of as the threshold. We go through the possible range of the threshold and determine the threshold to be 3.7.

4.2.6 To classify Species #6

We observe that the ratio $\frac{\text{PL}}{\text{HW}}$ differentiates Species #6 from the rest. Hence, we get a function shown below:

$$f(x) = \frac{x}{y} \tag{4.5}$$

Here x and y refer to the value of PL and HW. The following step is to determine the threshold for the result: when the result value of a lizard is larger than the threshold, it can be classified as Species #7. We go through the possible range of the threshold and determine the threshold to be 1.52.

4.2.7 To classify Species #8

We observe that the ratios $\frac{\text{SCGr}}{\text{HW}}$ differentiate Species #8 from the rest. Hence, we get a function shown below:

$$f(x) = \frac{x}{y} \tag{4.6}$$

Here x and y refer to the value of SCGr and HW. The following step is to determine the threshold for the result: when the result value of a lizard is smaller than the threshold, it can be classified as Species #8. We go through the possible range of the threshold and determine the threshold to be 0.95.

And the rest lizards are classified to Species #3.

4.3 Classification Result

Table 19: The result of species classification

Amount: True Species / Predict Species	#1	#2	#3	#4	#5	#6	#7	#8
#1	60	5	1	0	0	0	0	3
#2	5	54	4	0	0	0	0	2
#3	0	1	137	10	0	4	1	2
#4	0	0	1	81	0	0	0	0
#5	0	0	0	0	24	0	0	0
#6	0	2	11	1	0	113	1	1
#7	1	1	2	1	0	2	20	2
#8	0	0	0	0	0	1	0	10

From the table above, we can see that we mis-classify 63 lizards, which accounts for 11.5%. And the performance metric can then be calculated:

$$P_c = 85.83\%$$

5 Strengths and Weaknesses

5.1 Strengths

1. **Simplicity**

   The calculations that serve as part of the criterion were relatively straightforward and easy to complete, resulting in few errors. This in turn ensured that our models were both robust and sustainable, reducing the probability of errors in biological calculations leading to errors in judgment and also facilitating the work of biologists.

2. **Utility**

   It might be challenging to tell one animal species from another while examining it from the outside. Although ”key traits” are occasionally discovered (such as some unique coloring), scientists frequently have to rely on sets of quantifiable properties. As a result, our model is practical.

3. **Maneuverability**

   Our calculations for the criteria only use multiplication and division, which are simple to obtain on graphing calculators and can be calculated without one, demonstrating how flexible and operable the criteria we set are. The user experience is also excellent in the meantime.

5.2 Weaknesses

1. **Inaccuracy**

   Despite the fact that our model was quite successful at distinguishing Species #5, we were unable to discover features that could distinctly differentiate some species from the others, reducing the overall accuracy of the classification.

2. **Fragility**

   Only 564 lizards' data are provided, making the sample size relatively small. Some species have significantly insufficient sample sizes. As a result, the generalizability of our model's criterion is weak. In other words, our criteria might not be well-applicable if additional data is provided.

Reference

[1] Mariane U. V. Ronque, Marianne Azevedo-Silva, Gustavo M. Mori, Anete P. Souza, Paulo S. Oliveira, Three ways to distinguish species: using behavioural, ecological, and molecular data to tell apart two closely related ants, Camponotus renggeri and Camponotus rufipes (Hymenoptera: Formicidae), Zoological Journal of the Linnean Society, Volume 176, Issue 1, January 2016, Pages 170–181,