IMMC 2023 Problem E (Greater China, Winter) (English 简体 繁體)

How to distinguish biological species by numbers?

Many closely related animal species are hard to distinguish from each other by their exterior appearance. Sometimes “key features” may be found (e.g., some special colouring), but often the biologists have to rely upon sets of measurable characteristics.

You are provided with a set of real data containing the measurements of 564 lizards of 8 species belonging to genus *Darevskia*. The measurements consist of the counts of scales on different parts of thelizards’ bodies (pholidosis characteristics) and linear sizes of the lizards’ body parts (morphometric characteristics). For each lizard, its number-coded biological species and sex are given.

You are required to develop criteria that enable one to predict biological species and sex of the lizards with the best possible accuracy on the basis of such measurements. These criteria must be relatively simple and obvious, i.e. be realistically calculatable by a biologist in field conditions using nothing more than an graphing calculator. Solution of the computational “black box” type (e.g., artificial neural network classifiers or any other “closed” program that does not explain its “internal logic”) are not suggested. It is simple and obvious criteria that may enable the biologists to construct explanatory theories.

Tasks

1) Build a criterion that, with the best possible accuracy, differentiates lizards of species #5 from allother lizards and uses only the femoral pore number on the right side (FPNr). You may plot and explore the FPNr distribution of the lizards with regard to their species.

2) Build a criterion that, with the best possible accuracy, differentiates lizards of species #5 from allother lizards and uses two variables out of the measured morphometric and pholidosis characteristics. One of the methods to find the best pair of predicting variables (predictors) may beexhaustive search over all possible pairs of variables.

3) Build a criterion that, with the best possible accuracy, predicts lizards’sex regardless of their biological species on the basis of the morphometric and/or pholidosis characteristics. It is expected (but not guaranteed!) that sex is correlated with the ratiosof some measured linear sizes; but this does preclude from using other predictors in the criterion.

4) Not all lizard species in question are found in the same locations. Therefore, in practice, the tasks of distinguishing species from certain subgroups that live together are most often encountered. Build a set of criteria that, with the best possible accuracy, differentiate all species within the following groups:
    a. species #6 and #7,
    b. species #1 and #2,
    c. species #3, #4, and #5.

5) Build a criterion or a set of criteria that, with the best possible accuracy, predict lizards’ species or species and sex in their entire population (this may be useful to the biologists if they don’t know the locality where the lizard was captured).

It is quite possible that some pairs or groups of species will be inseparable on the basis of the available data. Provide the best obtained result that may be most helpful to the biologists.

General requirement. All your criteria must be accompanied by their performance metrics, i.e. the numbers of correctly and incorrectly classified lizards (preferably with a breakdown by true classes). For example, in Task 1, the "full" metrics is contained in the following table (where you have to fill cells *a*, *b*,*c*, and *d* with numbers):

	True species #5	True species #1-4, 6-8
Classified as species #5	*a*	*b*
Classified as species #1-4, 6-8	*c*	*d*

Correctly classified are *a* and *d*; incorrectly classified (classification errors) are *b* and *c*.

Remark 1. Neither of the Tasks is strictly mandatory. Although the higher number of successfully solved tasks improves the work's performance.

Remark 2. The Tasks are ordered by their expected difficulty. If you create a method to solve a "hard"Task, it will likely enable you to solve all previous Tasks with little effort.

Remark 3. The problem is a real applied research problem with real data. Thus, the "perfect" solutionmay not exist at all. In that case, the "best" solution is the "least bad" one.

Remark 4. An example of a "simple criterion": an explicitly specified function *f* of one or several measured variables $p1$, $p2$, … and a condition like "if $f(p1, p2, \ldots) > h$, then the lizard belongs to that class, otherwise it belongs to another class". Function *f* must be "calculatable on an graphing calculator", i.e. it may contain "standard" functions (power functions, trigonometric functions, exponents, logarithms, etc.), but it cannot contain a hidden iterative algorithm or calculations in volume beyond the capabilities of manual implementation.

Data description

The supplied XLSX-files contain the following columns:

Species_num – number-coded species, integer from 1 to 8;

Sex_num – number-coded sex, 1=male, 2=female;

Sex – letter-coded sex, M=male, F=female;

All other columns are the lizards' characteristics and are described below:

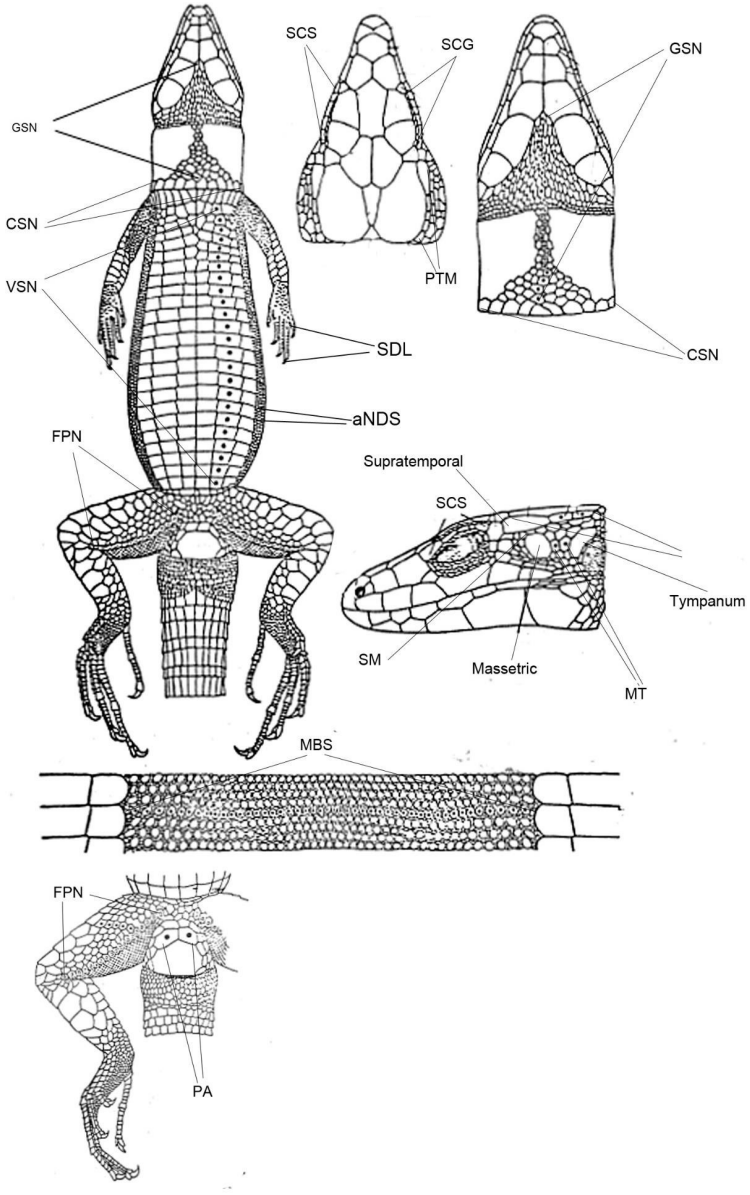
Pholidosis characteristics (counts of scales):

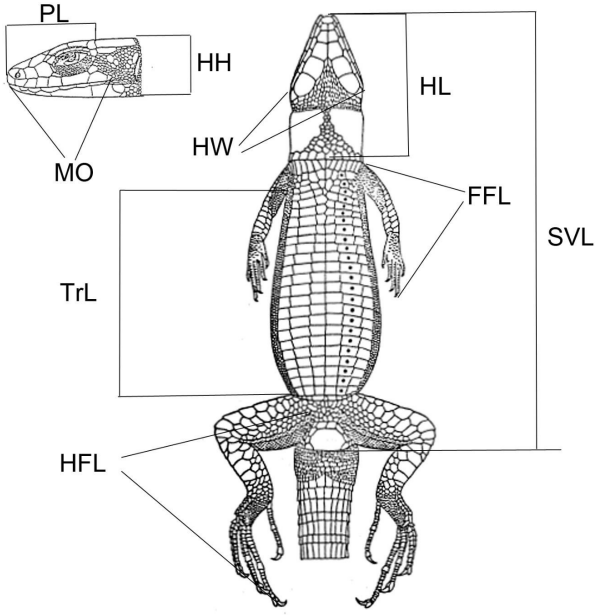
1. MBS - medium body scales, number of dorsal scales, approximately at half trunk;
2. VSN - ventral scale number on the middle line;
3. CSN - collar scale number;
4. GSN - gular scale number from the angle between the maxillar scales to the collar;
5. FPN - femoral pore number (FPNr – FPN on the right side);
6. SDL - subdigital lamellae in the 4th toe of the forelimb (SDLr –SDL on the right forelimb);
7. SCS - number of superciliary scales (SCSr – SCS on the right);
8. SCG - number of superciliary granules (SCGr – SCG on the right);
9. SM - number of scales between the masseteric shield and the supratemporal scale (SMr – SM on the right);
10. MT - number of scales between masseteric and tympanum shields on the right (MTr – MT on the right);
11. PA - preanal scale number;
12. PTM - postemporal scale number (PTMr – PTM on the right);

13. aNDS – average number of dorsal scales along one abdomen scale near limb (aNDSr – aNDS on the right);

Morphometric characteristics (all lengths are in millimeters):

14. SVL - snout-vent length, length of the body from tip of snout to cloaca;
15. TRL - trunk length (from the groin to the armpit);
16. HL - head length, measured ventrally from the tip of the snout to the posterior margin of the collar;
17. PL - pileus length measured dorsally from the tip of the snout to the posterior margin of the parietal + occipital scales;
18. ESD - length of the posterior half of the pileus, measured from the anterior margin of the 3rd supraocular scale to the posterior margin of the parietal + occipital scales;
19. HW - width of the head before the tympanic hole;
20. HH - head height near the occipital plate;
21. MO - mouth opening, measured laterally from the tip of the snout to the end of the mouth;
22. FFL - total forelimb length, from the base to the tip of the longest toe;
23. HFL - total hindlimb length, from the base to the tip of the longest toe.





Submission

Your solution paper should include a 1-page Summary Sheet. The body cannot exceed 20 pages for a maximum of 21 pages with the Summary Sheet inclusive. The appendices and references should appear at the end of the paper and do not count towards the 21 pages limit.

如何通过数字区分生物种类？

许多近缘动物的种/物种的类别很难通过外观相互区分。有时可能会发现“关键特征”（例如，一些特殊的色彩），但生物学家通常须依赖可测量特征的数据集。

您将获得一组真实数据，其中包含 Darevskia 属 8 个种/物种的 564 只蜥蜴的测量值。测量值包括蜥蜴身体不同部位的鳞片计数（鳞序特征）和蜥蜴身体部位的线性大小（形态学特征）。对于每只蜥蜴，都给出了其数字编码的生物物种和性别。

您需要开发标准，使人们能够在这些测量值的基础上尽可能准确地预测蜥蜴的生物种类和性别。这些标准须相对简单明了，即可以让生物学家在野外条件下仅使用图形计算器就可做实际的计算。不建议采用计算“黑匣子”类的解决方案（例如，人工神经网络分类器或任何其他不解释其“内部逻辑”的“封闭”程序）。正是简单明了的标准可以有助生物学家构建解释性理论。

任务

1) 建立一个标准，尽可能准确地将第 5 种蜥蜴与所有其他蜥蜴区分开来，并仅使用右侧的股骨孔数 (FPNr)。您可以绘制并探究蜥蜴在其物种方面的 FPNr 分布。

2) 建立一个标准，尽可能准确地将第 5 种蜥蜴与所有其他蜥蜴区分开来，并使用测量的形态学和鳞序特征中的两个变量。找到最佳“预测变量对”（预测变量）的方法之一可能是对所有可能的变量对做穷尽搜索。

3) 建立一个标准，以尽可能准确的方式预测蜥蜴的性别，而不考虑它们的生物物种，基于形态学和/或鳞序特征。生物学家预计（但不能保证！）性别与某些测量的线性大小的比率相关；但这确实排除了在标准中使用其他预测变量。

4)问题中并非所有的蜥蜴物种都出现在同一地点。因此，在实践中，最常遇到的任务是将物种与某些生活在一起的亚种群区分开来。 建立一套标准，尽可能准确地区分以下组内的所有物种：

a. 物种 #6 和 #7，
b. 物种 #1 和 #2，
c. 物种 #3、#4 和 #5。

5) 建立一个标准或一组标准，以尽可能准确的方式预测蜥蜴在整个种群中的种类和性别（如果生物学家不知道蜥蜴的捕获地，这可能对他们很有帮助）。

根据现有数据，某些物种对或种群很可能是不可分割的。请提供对生物学家最有帮助的最佳结果。

一般要求。您的所有标准都必须附有它们的性能指标，即正确和错误分类的蜥蜴的数量（最好按真实类别进行细分）。例如，在任务 1 中，“完整”指标包含在下表中（您必须在其中用数字填充单元格 a、b、c 和 d）：

	真实物种 #5	真实物种#1-4, 6-8
被划分为物种 #5	*a*	*b*
被划分为物种 #1-4, 6-8	*c*	*d*

正确分类的是 a 和 d； 错误分类（classification errors）的是 b 和 c。

说明 1. 每项任务并非必须完成，尽管成功解决的任务数量越多，您的作品表现就会越好。

说明 2. 任务是按难度排序。如果您创建一种方法来解决“困难”任务，它很可能使您能够毫不费力地解决所有以前的任务。

说明 3. 本题目是基于真实数据的真实的应用研究问题。因此，“完美”的解决方案可能压根不存在。在这种情况下，“最好”的解决方案是“最不坏”的解决方案。

说明 4.“简单标准”的示例：一个或多个测量变量 p1、p2、… 的显式表达函数 f 和条件，如“如果 f(p1、p2、…)>h，则蜥蜴属于那个种类，否则它属于另一个种类”。函数 f 必须是“可在图形计算器上计算的”，即它可以包含“标准”函数（幂函数、三角函数、指数、对数等），但不能包含隐藏的迭代算法或超出手动能力范围的计算量。

数据说明

本题提供的 XLSX 文件包含以下数据列：

Species_num - 数字编码的物种，从 1 到 8 的整数；
Sex_num - 数字编码的性别，1=雄性，2=雌性；
Sex -字母编码的性别，M=雄性，F=雌性；

所有其他数据列都是蜥蜴的特征，如下所述：

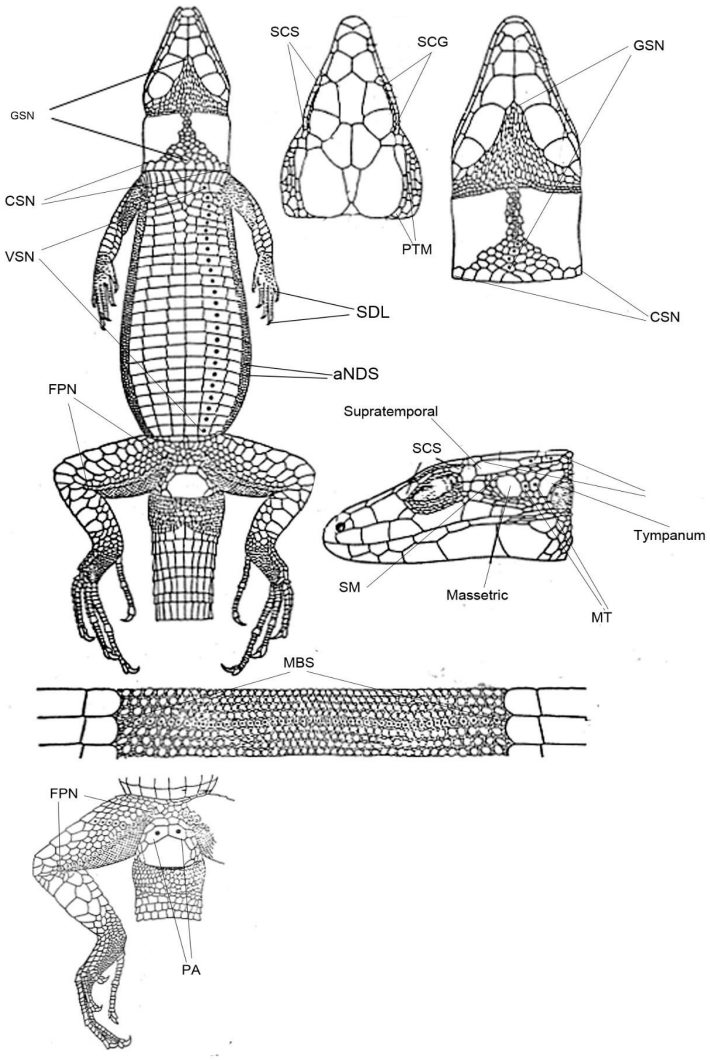
鳞序特征（鳞片数）：

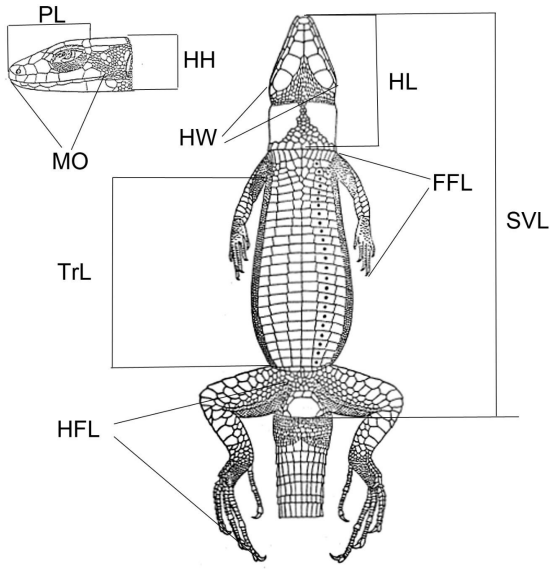
1. MBS——体鳞中等，背鳞数量，大约在躯干的一半处；
2. VSN——中线上的腹鳞数；
3. CSN——领标号；
4. GSN——从上颌鳞片到领口的角的角鳞片数；
5. FPN——股骨孔数（FPNr——右边的 FPN）；
6. SDL - 前肢第 4 个脚趾的趾下薄片（右前肢的 SDLr -SDL）；
7. SCS - 眉上鳞片数量（右侧的 SCSr - SCS）；

8. SCG – 眉上颗粒数（SCGr – 右侧的 SCG）；
9. SM – 咬肌盾和颞上鳞片之间的鳞片数量（右侧的 SMr – SM）；
10. MT——右边咬肌盾和鼓室盾之间的鳞片数量（MTr——右边的 MT）；
11. PA——肛前鳞数；
12. PTM——后时间鳞数（PTMr——右边的 PTM）；
13. aNDS – 靠近四肢的一个腹部鳞片上的平均背鳞数（aNDSr – 右侧的 AnDS）；

形态特征（所有长度均以毫米为单位）：

14. SVL – 鼻孔长度，从鼻尖到泄殖腔的身体长度；
15. TRL——躯干长度（从腹股沟到腋窝）；
16. HL – 头长，从鼻尖到项圈后缘的腹侧测量；
17. PL – 从鼻尖到顶叶 + 枕骨鳞片后缘的背面测量的菌盖长度；
18. ESD – 伞状盖后半部的长度，从第 3 眼上鳞片的前缘到顶叶 + 枕骨鳞片的后缘测量；
19. HW——鼓室前头部宽度；
20. HH——靠近枕骨板的头部高度；
21. MO – 嘴巴张开度，从鼻子尖端到嘴巴末端横向测量；
22. FFL – 总前肢长度，从基部到最长脚趾的尖端；
23. HFL – 后肢总长度，从基部到最长脚趾的尖端。





提交

您团队的解决方案论文应包括 1 页的摘要。正文不能超过 20 页，含摘要最多 21 页。附录和参考资料应出现在正文之后，不算在 21 页的限制之内。

如何通過數字區分生物種類？

許多近緣動物的種/物種的類別很難通過外觀相互區分。有時可能會發現「關鍵特徵」（例如，一些特殊的色彩），但生物學家通常須依賴可測量特徵的數據集。

您將獲得一組真實數據，其中包含 Darevskia 屬 8 個種/物種的 564 只蜥蜴的測量值。測量值包括蜥蜴身體不同部位的鱗片計數（鱗序特徵）和蜥蜴身體部位的線性大小（形態學特徵）。 對於每隻蜥蜴，都給出了其數位編碼的生物物種和性別。

您需要開發標准，使人們能夠在這些測量值的基礎上儘可能準確地預測蜥蜴的生物種類和性別。 這些標準須相對簡單明瞭，即可以讓生物學家在野外條件下僅使用圖形計算機就可做實際的計算。不建議採用計算「黑匣子」類的解決方案（例如，人工神經網路分類器或任何其他不解釋其內部邏輯的「封閉」 程式）。正是簡單明瞭的標準可以有助生物學家構建解釋性理論。

任務

1) 建立一個標準，盡可能準確地將第 5 種蜥蜴與所有其他蜥蜴區分開來，並僅使用右側的股骨孔數（FPNr）。您可以繪製並探究蜥蜴在其物種方面的 FPNr 分佈。

2) 建立一個標準，盡可能準確地將第 5 種蜥蜴與所有其他蜥蜴區分開來，並使用測量的形態學和鱗序特徵中的兩個變數。找到最佳「預測變數對」（預測變數）的方法之一可能是對所有可能的變數對做窮盡搜索。

3) 建立一個標準，以盡可能準確的方式預測蜥蜴的性別，而不考慮它們的生物物種，基於形態學和/或鱗序特徵。 生物學家預計（但不能保證！）性別與某些測量的線性大小的比率相關；但這確實排除了在標準中使用其他預測變數。

4)問題中並非所有的蜥蜴物種都出現在同一地點。因此，在實踐中，最常遇到的任務是將物種與某些生活在一起的亞種群區分開來。建立一套標準，盡可能準確地區分以下組內的所有物種：

a. 物種 #6 和 #7，
b. 物種 #1 和 #2，
c. 物種 #3、#4 和 #5。

5) 建立一個標準或一組標準，以盡可能準確的方式預測蜥蜴在整個種群中的種類和性別（如果生物學家不知道蜥蜴的捕獲地，這可能對他們很有幫助）。

根據現有數據，某些物種對或種群很可能是不可分割的。 請提供對生物學家最有說明的最佳結果。

一般要求。您的所有標準都必須附有它們的性能指標，即正確和錯誤分類的蜥蜴的數量（最好按真實類別進行細分）。例如，在任務 1 中，「完整」指標包含在下表中（您必須在其中用數位填充單元格 a、b、c 和 d）：

	真實物種 #5	真實物種#1-4, 6-8
被劃分為物種 #5	*a*	*b*
被劃分為物種 #1-4, 6-8	*c*	*d*

正確分類的是 a 和 d； 錯誤分類的是 b 和 c。

說明 1. 每項任務並非必須完成，儘管成功解決的任務數量越多，您的作品表現就會越好。

說明 2. 任務是按難度排序。 如果您創建一種方法來解決「困難」任務，它很可能使您能夠毫不費力地解決所有以前的任務。

說明 3. 本題目是基於真實數據的真實的應用研究問題。 因此，「完美」的解決方案可能壓根不存在。 在這種情況下，「最好」的解決方案是「最不壞」的解決方案。

說明 4. 「簡單標準」的示例：一個或多個測量變數 p1、p2、... 的顯式表達函數 f 和條件，如「如果 f（p1、p2、...）>h，則蜥蜴屬於那個種類，否則它屬於另一個種類」。函數 f 必須是「可在圖形計算機上計算的」，即它可以包含「標準」函數（冪函數、三角函數、指數、對數等），但不能包含隱藏的反覆運算演演算法或超出手動能力範圍的計算量。

數據說明

本題提供的 XLSX 檔包含以下數據列：

Species_num - 數字編碼的物種，從 1 到 8 的整數；
Sex_num - 數字編碼的性別，1=雄性，2=雌性；
Sex - 字母編碼的性別，M=雄性，F=雌性；

所有其他數據列都是蜥蜴的特徵，如下所述：

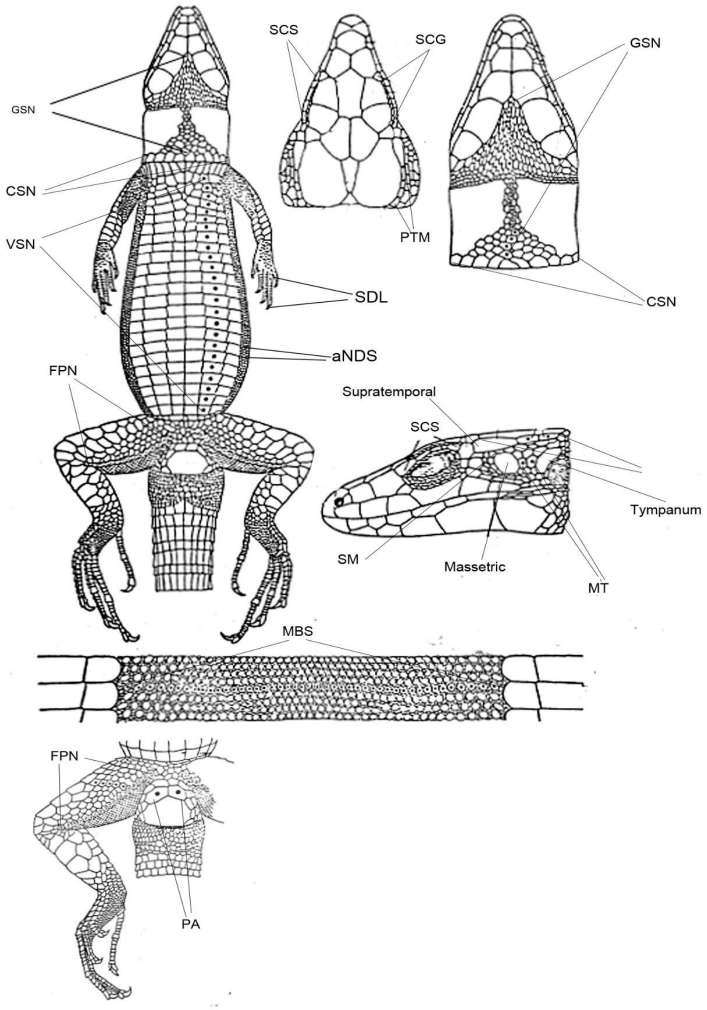
鱗序特徵（鱗片數）：

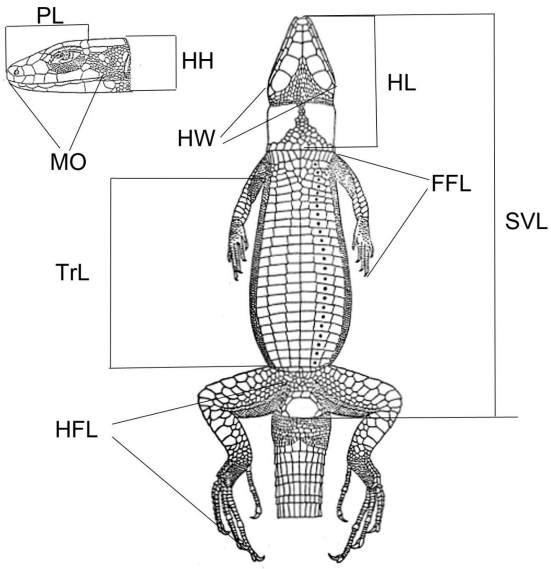
1. MBS——體鱗中等，背鱗數量，大約在軀幹的一半處；
2. VSN——中線上的腹鱗數；
3. CSN——領標號；
4. GSN——從上頜鱗片到領口的角的角鱗片數；
5. FPN——股骨孔數（FPNr——右边的 FPN）；
6. SDL - 前肢第 4 個腳趾的趾下薄片（右前肢的 SDLr - SDL）；
7. SCS - 眉上鱗片數量（右側的 SCSr - SCS）；

8. SCG – 眉上顆粒數（SCGr – 右側的 SCG）；
9. SM – 咬肌盾和顳上鱗片之間的鱗片數量（右側的 SMr – SM）；
10. MT –右邊咬肌盾和鼓室盾之間的鱗片數量（MTr –右邊的 MT）；
11. PA –肛前鱗數；
12. PTM –後時間鱗數（PTMr——右邊的 PTM）；
13. aNDS – 靠近四肢的一個腹部鱗片上的平均背鱗數（aNDSr – 右側的 AnDS）；

形態特徵（所有長度均以毫米為單位）：

14. SVL – 鼻孔長度，從鼻尖到泄殖腔的身體長度；
15. TRL – 軀幹長度（從腹股溝到腋窩）；
16. HL – 頭長，從鼻尖到項圈后緣的腹側測量；
17. PL – 從鼻尖到頂葉 + 枕骨鱗片后緣的背面測量的菌蓋長度；
18. ESD – 傘狀蓋後半部的長度，從第 3 眼上鱗片的前緣到頂葉 + 枕骨鱗片的后緣測量；
19. HW – 鼓室前頭部寬度；
20. HH – 靠近枕骨板的頭部高度；
21. MO – 嘴巴張開度，從鼻子尖端到嘴巴末端橫向測量；
22. FFL – 前肢總長度，從基部到最長腳趾的尖端；
23. HFL – 後肢總長度，從基部到最長腳趾的尖端。





提交

您團隊的解決方案論文應包括 1 頁的摘要。正文不能超過 20 頁，含摘要最多 21 頁。 附錄和參考資料應出現在正文之後，不算在 21 頁的限制之內。