Instructor: LCK & WJ  
TA: YJC, LFK, & YTL

Lab 0: DDM survey

The goal of this lab is to survey students’ prior knowledge and experience on the topic of DDM.

N.B. we are NOT looking for perfect answers; instead, we hope the Lab 0 can help us to seek a better way to work with most of students in the class.

N.B. You will still get good scores if you fail to come up with a conclusion but explain your thought process/concerns.

DUE: February 26 (Sunday), 2023 11:59 am (即將到來的這個星期日中午前)

Question 0 (0 pt)

1. Your name: 廖泓傑
2. Your department (若為資管/圖資雙主修視同本系同學，請特別註明): 財金雙主修資管
3. Your student ID: b07703001
4. 授權碼you are requesting : IM5053 or LIS5098 (please pick only one) IM5053
5. Your email (which you can receive 授權碼): b07703001@ntu.edu.tw

Question 1 (1 pt)

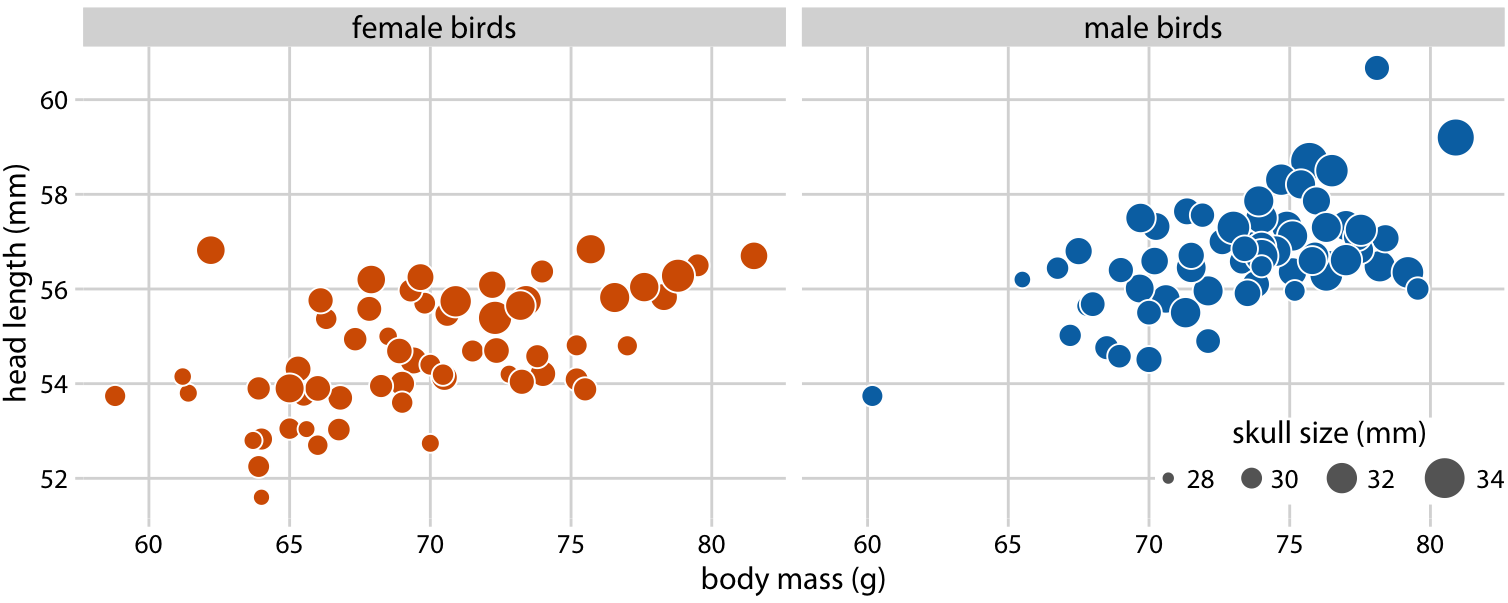


Figure 1. Head length versus body mass for 123 blue jays. (source: Wilke)

As shown in Figure 1, the birds’ sex is labeled by color and the birds’ skull size by circle size. Head-length measurements include the length of the bill (鳥喙), while skull-size measures do not. We found head length and skull size are potentially correlated; however, we also observed some birds with unusually long or short bills given their skull size. Data source: Keith Tarvin, Oberlin College

Please briefly comment on Fig 1:

* What did you learn from the figure?
* If one is asked to merge two sub-figures into one, is there anything you would like your audience to pay attention to?

Your Answers (中文ok):

1. **What did you learn from the figure?**

In general, the figures indicate that any two out of the three measurement of a blue jay, the skull size, body mass and head length, have positive correlations. Moreover, the correlations can be found in both genders. In most cases, birds that are heavier tends to have a higher average of skull size and head length, and vice versa. However, there is no guaranteed relationship between the three traits, there are exceptions that some individuals do not follow the tendency.

Comparing the two sub-graphs, it can be found that male birds generally are heavier, and have longer head length and bigger skull size. By observing birds with body mass between 65 ~75 g. It can also be noticed that given birds with the same body mass, male birds tend to have longer head length and bigger skull size than female birds. Again, the relationship is not guaranteed and there are exceptions that some individuals do not follow the observation.

1. **If one is asked to merge two sub-figures into one, is there anything you would like your audience to pay attention to?**

To merge the two sub-figures, I will continue to use colors to identify their sex to make comparing male and female birds visually more convenient. I would like my audience pay attention to the findings I mentioned in the first part, that male birds are generally bigger in all three traits compared to female birds, and given the same body mass, male birds tend to have longer head length and bigger skull.

As the passage suggests, we observed some birds with unusually long or short bills given their skull size, but the length of bills or not on the graph directly. In this case, I would replace the axis of head length by bill length to emphasize our finding to be more intuitive.

Question 2 (1 pt)

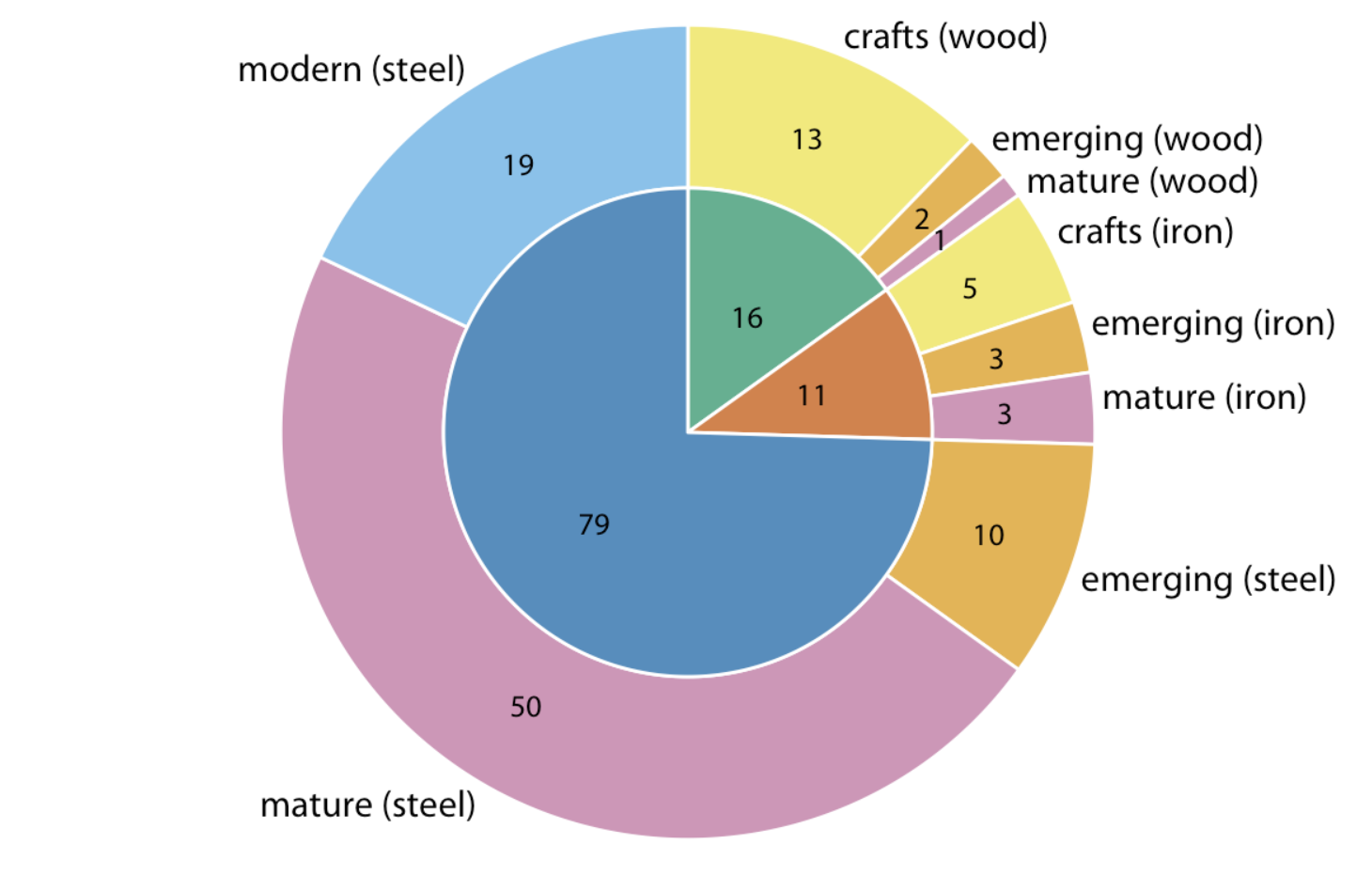


Figure 2. The breakdown of bridges in Pittsburgh in the US by construction material (Wilke)

Data source: Yoram Reich and Steven J. Fenves, via the UCI Machine Learning Repository (Dua and Karra Taniskidou 2017)

Figure 2 shows a visual representation of bridges in Pittsburgh with details on their construction materials (such as steel, wood, iron) and the era in which they were built (including crafts, emerging, mature, and modern periods). The numbers shown indicate how many bridges fall into each category.

Please briefly comment on Fig 2:

* What did you learn from the figure?
* Is there any other limitation or drawback that you can identify in Figure 2?

Your Answers (中文ok):

1. **What did you learn from the figure?**

By observing the figure, we can know that there are total 106 bridges in Pittsburgh included. Grouping by their construction materials, 79 of them are made of steel, which is approximately 75% of the bridges; 16 of them are made of wood, which is approximately 15% of the bridges; and 11 of them are made of iron, which is approximately 10% of the bridges. Grouping by the era being built, 18 of them are made in craft period, which is approximately 17% of the bridges; 15 of them are made in emerging period, which is approximately 14% of the bridges; 54 of them are made in mature period, which is approximately 51% of the bridges; and 19 of them are made in modern period, which is approximately 18% of the bridges.

If we consider the transformation of the materials being used in different era, newer bridges built in the modern and mature period are mostly built in steel. We can therefore assume that building a steel bridges is a newer technique being invented and being the mainstream of the two newer eras. Iron bridges can be found in craft, emerging, and mature periods, but it seems not being used in modern era. Wood bridges are mostly built in craft era, and is the first to be eliminated.

It can also be noticed that the era that built the most bridges is the mature era. While modern era, which should have the most advanced technology, does not share the same productivity. I assume that it may be either the total time length of the era not as long as the mature era, or the demand for bridges in Pittsburgh had gradually approached to saturation.

1. **Is there any other limitation or drawback that you can identify in Figure 2?**

While the graph is convenience for people to when a specific type of bridge is being built, it is not equally intuitive for people to know which type of bridge are preferred in certain period of time. Moreover, using two circles to represent the data is somehow a little confusing. I would prefer using conspicuous lines to separate building material and use color to identify which time it was built. Which can represent the same concept by using only one circle.

Additionally, the figure fails to specify the define the ranges of each era. I found it confusing when comparing the total amount of bridges being built in different periods. The situation could be caused by different time length of an era or other factors that are worth being discussed.

Question 3 (4 pts)

在雲端資料夾Lab 0的「hotel\_data.xlsx」這個檔案中，資料集包含美國某連鎖旅館品牌旗下100間旅館某一年的相關記錄。資料中 7 個欄位項目說明如下：

|  |  |
| --- | --- |
| 變數 | 意義 |
| 營業邊際利潤 | 營業邊際利潤，單位為百分點。這個數字一定介於0和100之間，數字愈大表示該旅館的營業績效愈好 |
| 附近客房總數 | 在該旅館 3 英哩內的其他旅館的客房總數，單位為間數 |
| 最近的競爭者與自己的距離 | 該旅館與最近之另一家旅館的距離，單位為英哩 |
| 週邊辦公空間 | 在該旅館周邊社區中的辦公大樓總樓地板面積，單位為千平方呎 |
| 大學註冊人數 | 在該旅館周邊社區中大專院校的註冊人數，單位為千人 |
| 家庭收入 | 在該旅館周邊社區之中等家庭的平均年收入，單位為千美元 |
| 離市中心的距離 | 該旅館到市中心的距離，單位為英哩 |

說明：您是這個品牌的負責人，想要瞭解是哪些因子造成了各旅館的營業邊際利潤有所差異，並且量化這些因子對營業邊際利潤的影響力。您的目標是找出顯然是有用的因子，例如：在檢討各旅館績效時可以更公允一些，未來要找地點開新旅館時可以有個參考。

針對上述任務，請試著用 Python、Excel 或任何你喜歡的程式語言或軟體建立迴歸模型，於本文件中貼上迴歸模型的結果截圖，並用你的統計知識和技術解讀你的迴歸模型，以使你的模型對旅館經營管理有幫助。

Your Answers and Screenshots (中文ok):

為了分析在該公司旗下旅館之營業邊際利潤，我們納入了六個可能影響其營業邊際利潤的因子，並試圖量化這些因子對營業邊際利潤的影響力。

首先我先對該六個影響因子與營業邊際利潤作複迴歸分析，得出以下結果:

1. R平方以及調整後R平方是用來衡量這些旅店樣本的營業績效受到該六個因子解釋的部分有多少，其中R平方和Adjusted R平方分別為0.525和0.494，可以粗略的表示這些旅館的營業邊際效率約有五成是受到這六個因素影響。



1. 透過回歸模型可算出平方和(SS)與均方和(MS)，並算出此模型的顯著值，在顯著水準α=0.05的情況下，其顯著值0.000<0.05，可以拒絕虛無假設，代表此模型並非常態分佈，可以解釋營業邊際利潤變化與其變數的關係。



1. 根據此回歸模型的結果，可以得出其線性回歸的公式為:

透過此方程式，對於新出現之樣本，在給定期參數數據的前提下，我們能漸進不偏的預測其營業邊際利潤的期望值。



1. 透過各參數的係數，可以知道各參數在增加1單位下，預期營業邊際利潤增加的量。可以觀察到除了附近客房總數和離市中心距離都是越低越好外，其他的參數都在增加時能帶給旅店更好的營業邊際利潤。
2. 透過各參數的P值，我們可以知道該數據對營業邊際利率的影響是否顯著，在顯著水準α=0.05的前提下，對於P值大於0.05的兩個參數—「大學註冊人數」和「離市中心距離」，我們認為它對營業邊際利潤並沒有足夠顯著的影響，證明其非隨機變數，亦即其可能與營業邊際利潤無關。

綜合以上，我們可以透過第3點得出的回歸模型，作為未來我們對新旅館投資的參考，並需特別關注其附近客房總數、最近競爭者與自己的距離、周邊辦公空間和家庭收入的狀況，一方面可以給現存旅館對於其可改變現況的參考，一方面則可在未來愈發展的旅館建造方案做選擇的參考。

Question 4 (0 pt)

How familiar are you with these technologies/concepts related to DDM?

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | How familiar are you with these technologies or concepts related to DDM? | | | | | Are you interested? | |
|  | Never heard of | Slightly | Moderately | Very | Extremely | Quite a lot | Not very |
| Descriptive statistics |  |  |  | ✓ |  | ✓ |  |
| Inferential statistics |  |  | ✓ |  |  | ✓ |  |
| Probability |  |  |  | ✓ |  | ✓ |  |
| Python |  |  | ✓ |  |  | ✓ |  |
| Machine learning |  | ✓ |  |  |  | ✓ |  |
| Visualization principles |  | ✓ |  |  |  | ✓ |  |
| Visualization software such as Tableau |  |  | ✓ |  |  | ✓ |  |
| Data cleansing |  | ✓ |  |  |  |  | ✓ |

Question 5 (0 pt)

**Why do you want to take this course?**

I am quite interested in data science and machine learning, and I believe by learning the knowledge and techniques being taught in this class can be extremely useful in many areas. I am very looking forward to learn in this course

**How many hours do you plan to spend on this course each week?**

5~7 hours including lectures.

**Any other comments about this course?**

Not really, since I had only started the class for only 3 hours now.

上傳提醒：

#繳交作業時，請將本文件(.docx)與 .py、.ipynb或 .xlsx 檔壓縮成一個 zip檔於Google 表單上傳。為了顧及連假期間TA們的肝，且讓同學盡快拿到授權碼，請同學們上傳表單前再三檢查（如以其他裝置測試看看能不能順利開啟），無法順利開啟的 zip 檔案夾， 我們不另行通知補件。

#之後每週的作業量皆與 Lab0相似（根據以往同學回報：因人而異2-10小時），同學在寫Lab0的時候可順便評估一下自身loading。機器學習的週次，作業量會略多於Lab 0。