Instructor: LCK & WJ

Lab 1: Numeracy and Descriptive Stats

B07703001 廖泓傑

Due: 11:59 AM — 3/11 (Sat), 2023
Point: 5

中英文作答皆可

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

Task 1- Looking for Benford's Law in Taiwan (2 points)

For this task, please find a real-world dataset that is related to local events or issues and examine whether it follows Benford's Law. Even if your results do not conform to the law, we still encourage you to report them. It is important to note that Benford's Law is not considered scientific, but rather a heuristic. Therefore, we **equally value unsuccessful results** as much as successful ones.

ATTN: To maximize our learning opportunities, please try hard to avoid the cliché ones that everyone can relate (e.g., population, ballot), or cases we already discussed about in the class.

a. Names for your teammate: (If you teamed up with someone, **it's ok for you all to submit the same content**.)

R10126032 圖資碩二 丁昱寧

b. The case/dataset you would like to examine. Please briefly describe what the dataset is about and where people can find it. (aboutness/data source)

主題：102 年～112 年 1 月每月查緝的毒品數量（公克重）

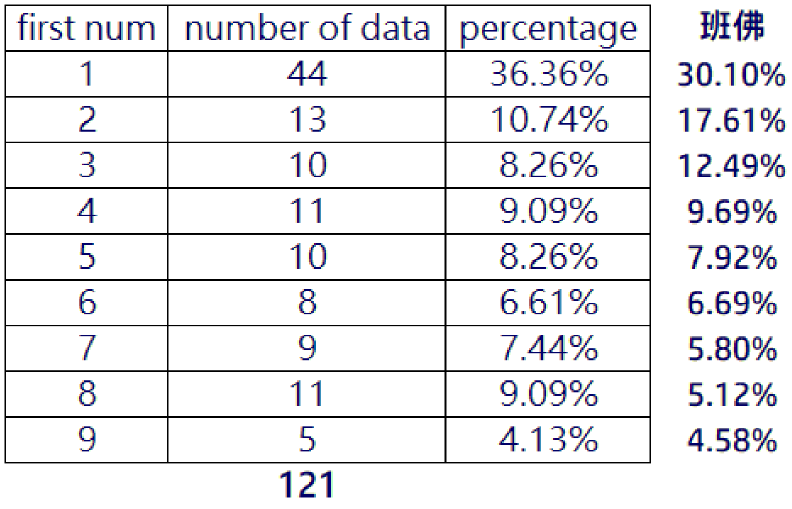
資料來源：內政部統計查詢網 https://statis.moi.gov.tw/micst/stmain.jsp?sys=100

c. Explain your rationales and observation— Why do you think it is worthwhile to test for compliance with Benford's Law?

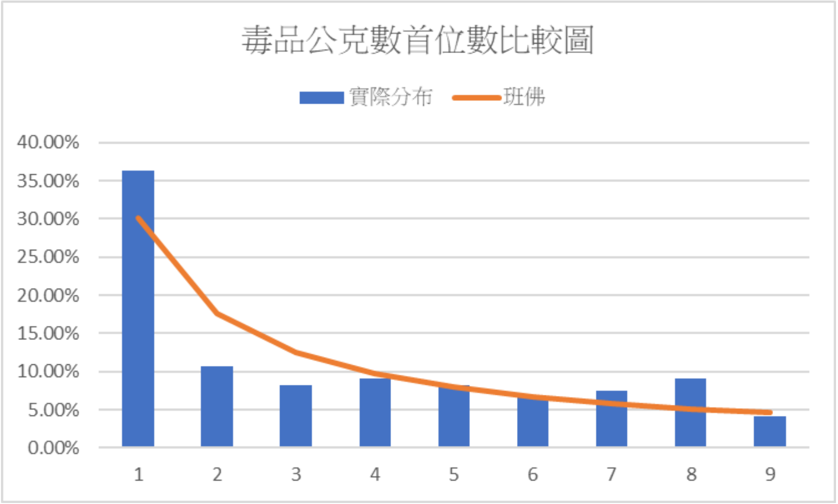
根據初步觀察，我國每月查緝的毒品數量高達數十萬甚至上百萬公克，資料數量大，各月份之數據差距大且分布隨機，故推測有機會符合班佛定律。

d. Findings— What were your findings? Did the dataset conform to Benford's Law? How did it conform, or not conform? Were you surprised by your results?

分析結果大致符合班佛定律。資料涵蓋從 102 年～112 年 1 月共 121 個月的 121 筆資料，當中毒品數量為 1 開頭的月份共有 44 個，佔了 36%，2 開頭佔 10.74%，3 開頭佔 8.26%......。詳細數據如下表所示:



first num	number of data	percentage	班佛
1	44	36.36%	30.10%
2	13	10.74%	17.61%
3	10	8.26%	12.49%
4	11	9.09%	9.69%
5	10	8.26%	7.92%
6	8	6.61%	6.69%
7	9	7.44%	5.80%
8	11	9.09%	5.12%
9	5	4.13%	4.58%
	121		



和班佛定律比較後，可發現除 1,2,3,8 有較明顯誤差，其餘大致符合班佛定律之分布。結果雖然如我們預期，但仍使人驚訝，尤其是在樣本數不到龐大的前提下仍能觀察到如此接近的分布，這種趨勢可以藉由班佛如此簡單的公式表示出來，讓人很佩服他的觀察力。

e. Have you shared your findings on COOL? y/n

Yes.

Task 2- Scales in cultural /social dimensions (1 point)

In this week's class, we learned about the different scales of measurement that are used to quantify data. However, some examples can be challenging to classify and have cultural or social dimensions to consider. Please select your preferred scale for each of the following examples and explain why we should always be careful about their scales when doing analysis.

a. Education levels on a questionnaire.
我認為教育程度應為次序尺度。因其為離散變數，且除了少數特例外，大部分情況須完成前一階段的教育才能進入下一階段，作為統計量時，教育程度也有高低之分。並且，在任兩個教育程度之間，彼此的差距並不固定，故為次序尺度。

b. Zip codes (i.e., postal codes) in Taiwan.
我認為郵遞區號為名目尺度。雖然郵遞區號由數字表示，但數字之間無高低之分，僅代表該地址位於哪個區域，故為名目尺度。

c. Peerage of the United Kingdom

我認為英國的 **peerage** 制度為次序尺度。因為總共只有五種爵位，且彼此間身分地位的距離難以測量，但這五種爵位的身分高低有絕對的關係，故為次序尺度。

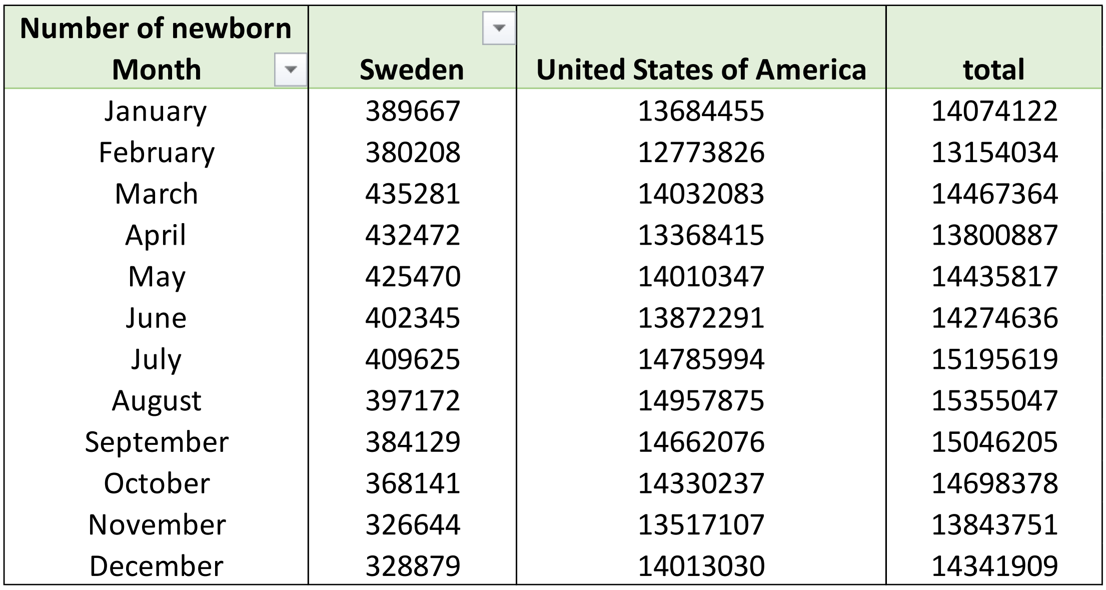
ATTN: For the grading, we place more emphasis on your reasoning and insights rather than which scale you choose. So, do not worry about taking a side if the options are confusing.

Task 3 - Cross tabulation (with a Pivot table) (1 point)

Dataset: Lab1-task3-dataset.xlsx
Description: The dataset presents the number of newborns in Sweden and the US, recorded by months and years.

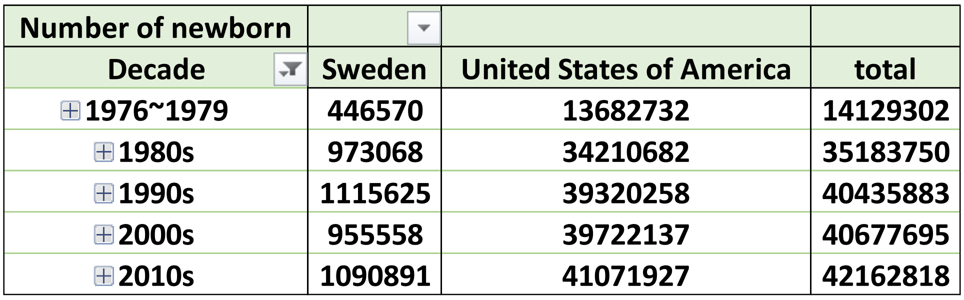
Please finish the following requests:

a. Generate a cross-tabulation table to show the number of newborns and the months of birth in both Sweden and the US.



Number of newborn Month	Sweden	United States of America	total
January	389667	13684455	14074122
February	380208	12773826	13154034
March	435281	14032083	14467364
April	432472	13368415	13800887
May	425470	14010347	14435817
June	402345	13872291	14274636
July	409625	14785994	15195619
August	397172	14957875	15355047
September	384129	14662076	15046205
October	368141	14330237	14698378
November	326644	13517107	13843751
December	328879	14013030	14341909

b. Generate a cross-tabulation table to show the number of newborns and the decades of birth (e.g., 1970s, 1980s) in both Sweden and the US.



Number of newborn			
Decade	Sweden	United States of America	total
⊞1976~1979	446570	13682732	14129302
⊞1980s	973068	34210682	35183750
⊞1990s	1115625	39320258	40435883
⊞2000s	955558	39722137	40677695
⊞2010s	1090891	41071927	42162818

c. Is there a seasonal variation in the number of births in both countries?
Yes, there is a seasonal variation in the number of births in both countries. In Sweden, the number of births is higher in March, April and May, while it is lower in November and December; In the USA, the number of births is higher in July, August and September, while it is lower in January and February. Although there are seasonal variation in both countries, the trends in the two countries are not identical.

Hints: Keep in mind that real-world datasets are often incomplete or imperfect. Please discuss how you handle incomplete data (if any) in this task.

For the number of births in 2020, the data in Sweden is included in the dataset but not the USA. Since it is not comparable to other numbers which is the number of births in a decade in question(b), I do not include the data regarding the number of births in Sweden in 2020. Also, the dataset starts recording from 1976, so I labeled 1976~1979 in question(b) to show the number while indicate that it is not comparable to other complete decades.

Task 4 - Descriptive stats (1 point)

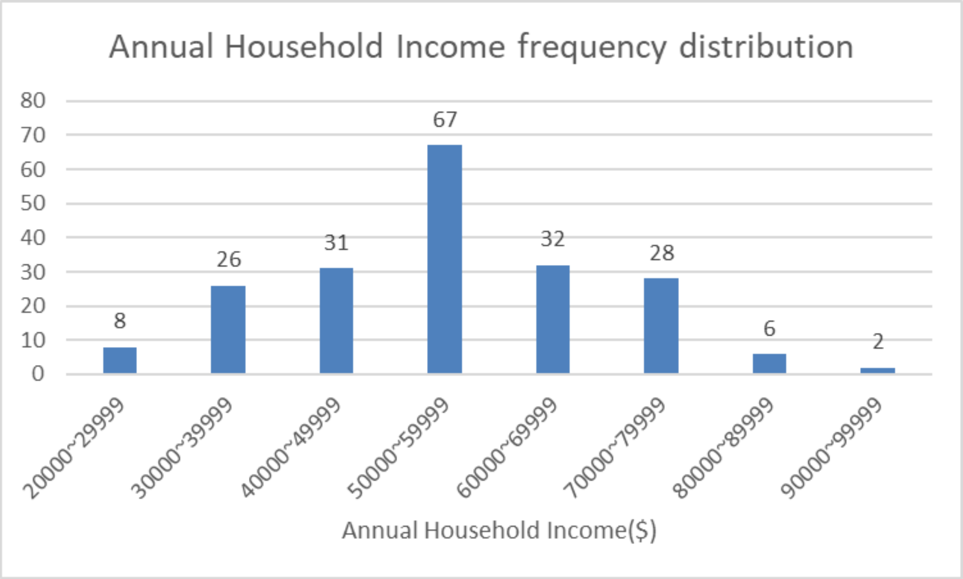
Dataset download: Lab1-task4-dataset.xlsx

The dataset contains data from 200 households randomly selected in the US. Columns A, B, and C contain the amounts spent on food in a year, the amount earned in a year, and the amount of non-mortgage debts. Column D records the region that each household belongs to, where 1 is for northeast, 2 is for midwest, 3 is for south, and 4 is for west. Column E records the location of the household, where 1 means inside a metropolitan area and 2 means the opposite. Please answer the following questions based on the data.

a. Prepare a frequency distribution for the data of annual household income (data in Column B). Let the lower bound of the first class be $20,000. Use an equal class interval of $10,000 for all classes.

Interval	Upper bound	Frequency
	20000	0
20000~29999	30000	8
30000~39999	40000	26
40000~49999	50000	31
50000~59999	60000	67
60000~69999	70000	32
70000~79999	80000	28
80000~89999	90000	6
90000~99999	100000	2

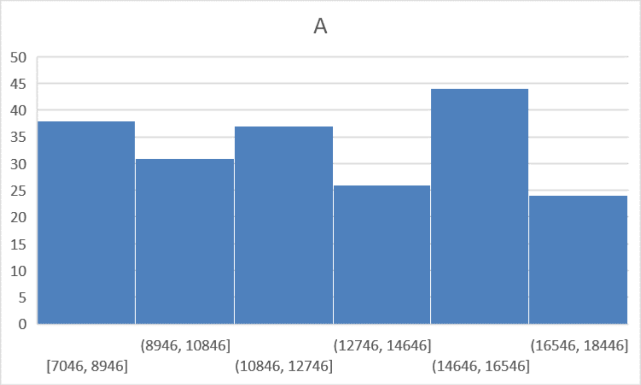
b. Generate a histogram showing the distribution (Column B). You may use any tool, but please remember to paste its screenshot.

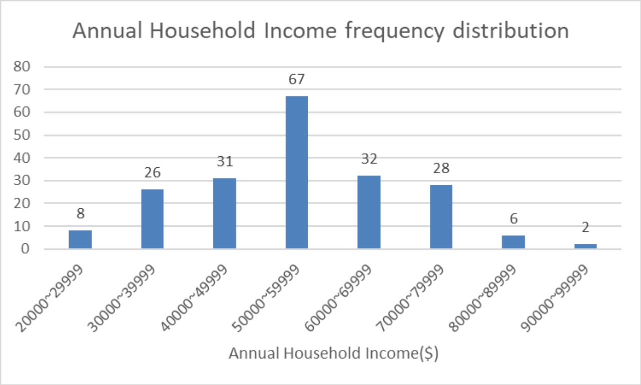


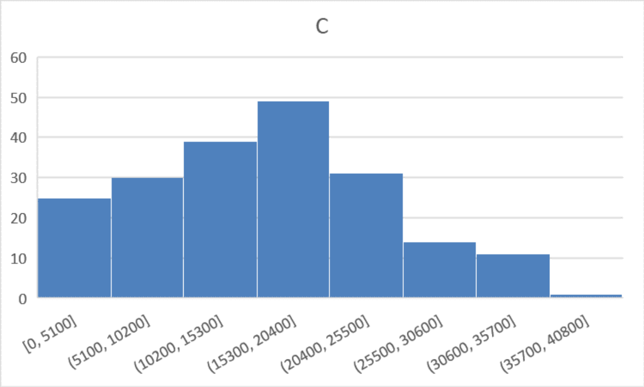
c. Based on the histogram, discuss the shape of the distribution.
根據此長條圖，可看出美國家庭年均收入的分布接近常態分布，在50000~59999此區間有最高的出現頻率(67筆資料)，並逐漸向兩邊遞減，且以50000~59999此區間為基準兩邊的資料分布比數很相像。

d. One of the three columns (A, B, or C) has fabricated data. Please identify which column you think it is and explain why.
觀察A、B、C三者的分布圖，可以發現在B(美國家庭年均收入)和C(非房貸負債)為接近常態分布的情況下，A(每戶食物支出)在各區間的分布相當平均。在

財富與花費這種較容易接近常態分布的樣本下，若三者有一者的數據為捏造的，我認為 A 數據最有可能是用經由亂數捏造的。







Task 5 -Install Tableau before Week 3 (0 points)

Before the week 3, please download and install the latest version of Tableau **“Desktop”** at https://www.tableau.com/products/desktop/download
Enter your NTU email address for “Business E-mail” and enter “National Taiwan University” for the Organization. Please access the **product key** at COOL.