TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO BÀI TOÁN 1

TP.Hồ Chí Minh, 4-2022

TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN

KHOA CÔNG NGHỆ THÔNG TIN

BÁO CÁO BÀI TOÁN 1

Sinh viên thực hiện: Lê Minh Trí - MSSV: 20120600

Lê Hữu Trọng - MSSV: 20120607

Lê Ngọc Đức - MSSV: 2012059

TP. Hồ Chí Minh, 4-2022

MỤC LỤC

1 Phần Trình Bày

1.1 Trả lời các yêu cầu:

- Xác định và hình thức hóa mục tiêu bài toán:
  - Câu hỏi cần giải quyết: Kiểm định giả thuyết: "Nhiệt độ trung bình từ năm 1978 đến năm 2018 **nhỏ hơn** nhiệt độ trung bình từ năm 1938 đến năm 1977 tại khu vực sân bay đảo Grand, bang Nebraska, Hoa Kỳ (Nghĩa là thời tiết có xu hướng trở lạnh từ năm 1978 đến 2018 tại khu vực này)"
  - Mô tả bài toán bằng ngôn ngữ toán học: Kiểm định giả thuyết hai mẫu, so sánh hai trung bình (trung bình nhiệt độ từ năm 1938 đến năm 1978 và nhiệt độ trung bình từ năm 1978 đến 2018) với phương sai chưa biết. Dạng toán này là kiểm định hai phía.

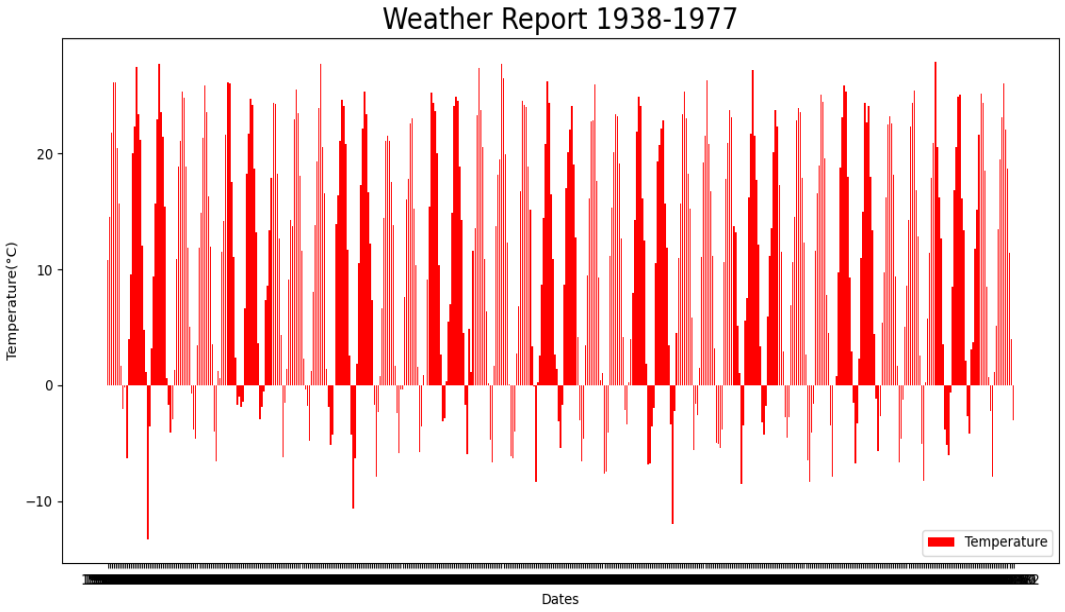
$$\begin{cases} H_0 : \mu_1 \leq \mu_2 \\ \mathrm{H}_1 : \mu_1 > \mu_2 \end{cases}$$

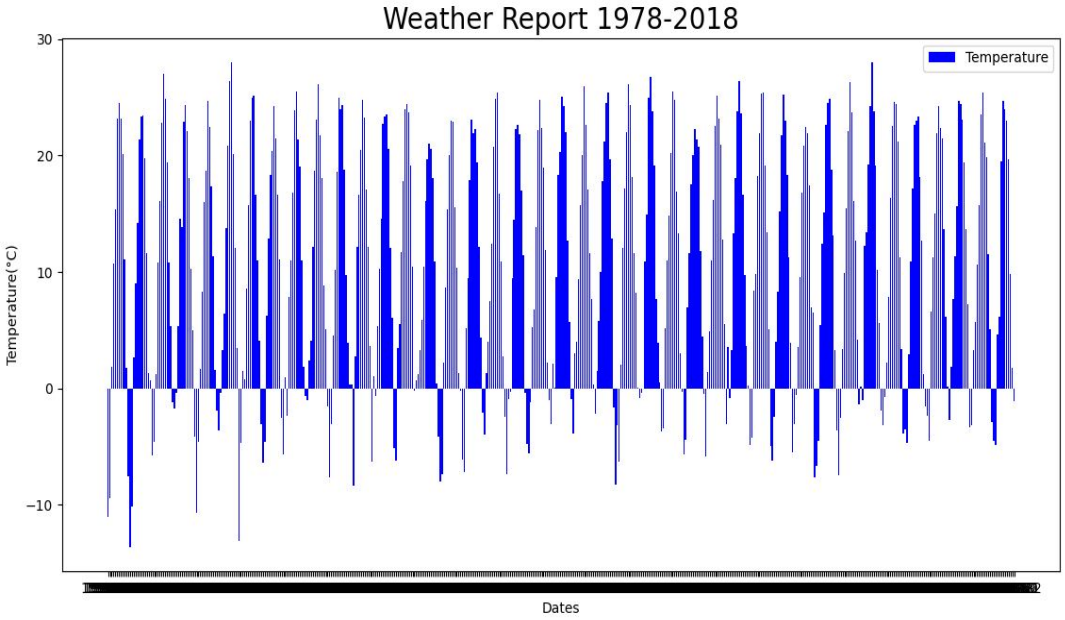
    Với μ_1 là nhiệt độ trung bình từ 1978 đến năm 2018, μ_2 là nhiệt độ trung ình từ 1938 dến 1977

- Dạng bài toán đưa ra: Suy luận (đặt ra giả thuyết và kiểm định giả thuyết đó)
- **Đối tượng dữ liệu sử dụng trong bài toán:** Nhiệt độ trung bình tính (Phạm vi thời gian) (Phạm vi không gian)
- Phạm vi, mức độ, quy mô của bài toán:
  - Theo thời gian: theo các tháng từ năm 1938 đến năm 2018
  - Theo không gian: tại khu vực sân bay đảo Grand, bang Nebraska, Hoa Kỳ

1.2 Thu thập và xử lý dữ liệu

Data visualization:





Nhận thấy được sự khác biệt về sự phân bố nhiệt độ giữa các tháng trong năm trong 2 tập dữ liệu (1938-1977 và 1978-2018). Từ 1938 đến 1977 có những tháng rất lạnh nhưng các tháng có nhiệt độ âm lại không phân bố đều. Từ năm 1978 đến 2018 không có những tháng cực lạnh như ở tập dữ liệu trước nhưng những tháng có nhiệt độ âm lại phân bố đều giữa các năm. Do đó em đưa ra dự đoán: "Nhiệt độ trung bình từ năm 1978

đến năm 2018 **nhỏ hơn** nhiệt độ trung bình từ năm 1938 đến năm 1977 tại khu vực sân bay đảo Grand, bang Nebraska, Hoa Kỳ (Nghĩa là thời tiết có xu hướng trở lạnh từ năm 1978 đến 2018 tại khu vực này)"

Dữ liệu được lấy tại `https://www.ncei.noaa.gov/data/global-summary-of-the-month/access/USW00014935.csv`

Ví dụ minh họa về dữ liệu:

U.S. Department of Commerce
National Oceanic & Atmospheric Administration
National Environmental Satellite, Data, and Information Service
Current Location: Elev: 1843 ft. Lat: 40.9615° N Lon: -98.3130° W
Station: **GRAND ISLAND CENTRAL NE REGIONAL AIRPORT, NE US USW00014935**

Global Summary of the Month for 1938
Generated on 04/02/2022

National Centers for Environmental Information
151 Patton Avenue
Asheville, North Carolina 28801

Date	Temperature (F)													Precipitation (Inches)									Observed Weather	
Elem ->	TAVG	TMAX	TMIN	HTDD	CLDD	EMXT		EMNT		DX90	DX32	DT32	DT00	PRCP	EMXP		SNOW	EMSD		DP01	DP10	DP1X	DYHF	DYTS
Month	Mean	Mean Max.	Mean Min	Heating Degree Days	Cooling Degree Days	Highest	High Date	Lowest	Low Date	Number of Days				Total	Greatest Observed		Snow, Sleet			Number of Days				
										Max >= 90	Max <= 32	Min <= 32	Min <= 0		Amount	Date	Total Fall	Max Depth	Max Date	>=.01	>=.10	>=1.0	FG+	TS
Apr	51.5	63.6	39.4	418	13	86	14	18	02	0	2	9	0	4.00	1.17	16	2.3	2	08	12	7	1		2
May	58.2	68.6	47.7	233	21	83	26	31	08	0	0	2	0	4.53	0.96	06	0.0	0	31	17	9	0		7
Jun	71.3	83.7	58.8	14	202	100	30	48	07	9	0	0	0	3.39	2.02	21	0.0	0	30	12	6	1		12
Jul	79.1	93.7	64.5	0	438	109	12	59	08	21	0	0	0	2.64	0.68	06	0.0	0	31	8	5	0		9
Aug	79.1	93.6	64.6	0	437	107	22	53	21	21	0	0	0	1.16	0.80	17	0.0	0	31	6	2	0		7
Sep	68.8	83.0	54.6	61	175	97	09	32	19	10	0	1	0	2.65	1.81	13	0.0	0	30	4	3	1		3
Oct	60.2	76.3	44.1	221	71	95	04	18	24	5	0	5	0	0.18	0.11	10	0.0	0	31	2	1	0		2
Nov	35.1	49.9	20.3	899	1	74	02	-2	25	0	5	27	2	0.02	0.01	06	0.1	0	30	2	0	0		3
Dec	28.3	42.5	13.9	1139	0	60	03	-4	28	0	2	30	3	0.03	0.02	04	0.1	0	31	2	0	0		

Notes

(Blank) Data element not reported or missing.
X Monthly means or totals based on incomplete time series.
\+ Occurred on one or more previous dates during the month. The date in the Date field is the last day of occurrence.

A Accumulated amount.
T Trace Amount.

FG+ Heavy Fog
TS Thunderstorms

Xử lý dữ liệu

Dữ liệu được lưu theo tháng tồn tại một số tháng không có dữ liệu (bị khuyết) nhưng không quan trọng vì mình tính nhiệt độ trung bình dựa trên số lượng các tháng có dữ liệu và nhiệt độ trong các tháng đó. Có thể thấy có nhiều dữ liệu dư thừa không cần sử dụng đến cho bài toán đặt ra bên trên như các cột TMAX(Trung bình các nhiệt độ cao nhất của các ngày trong tháng), TMIN(Trung bình các nhiệt độ thấp nhất của các ngày trong tháng), HTDD (Heating degree days: Lượng nhiệt cần để sưởi ấm),vv... Cột cần chú ý duy nhất để giải quyết bài toàn đặt ra là **TAVG** (Nhiệt độ trung bình trong tháng). Vì vậy cần có bước tiền xử lý dữ liệu để cho ra một tập tin chỉ chứa thông tin của tháng-năm và nhiệt độ tương ứng của nó. Sau khi xử lý dữ liệu trong dataset bên trên ta được tập tin TAVG_Dataset ở thư mục data hoặc xem tại: `https://drive.google.com/file/d/1roQA_HKOabYuqUjHW6TNeyCJxcygdzHA/view?usp=sharing`

1.3 Phân tích, đánh giá và kết luận:

Mô hình sử dụng: Định lý kiểm định so sánh hai trung bình khi phương sai khác nhau chưa biết:

$$\{H_0 : \mu_1 \leq \mu_2$$
$$\mathrm{H}_1 : \mu_1 > \mu_2$$

Ta định nghĩa thống kê kiểm định có phân phối Student với giá trị quyết định kiểm định:

$$T_0 = \frac{X_1 - X_2}{\sqrt{(S_1^2/n_1 + S_2^2/n_2)}}$$

và bậc tự do:

$$\nu = \left\lfloor \frac{\left(S_1^2/n_1 + S_2^2/n_2\right)^2}{\frac{(S_1^2/n_1)^2}{n_1 - 1} + \frac{(S_2^2/n_2)^2}{n_2 - 1}} \right\rfloor.$$

Với $\mu = \frac{1}{n} \sum \mathrm{a}_i$ và $s^2 = \frac{1}{n-1} \sum (a_i - \mu)^2$

Bác bỏ giả thuyết H_0 khi $T_0 \geq \mathrm{t}_{0.05,\nu}$ Chọn 0.05 vì mặc định mức ý nghĩa là 0.1. Ngược lại thì không đủ điều kiện bác bỏ H_0, tức là H_0 là đúng.

Các kết quả thu được:

Ký hiệu	Giải thích	Số liệu tính được
μ_1	Nhiệt độ trung bình từ năm 1978-2018	10.467
μ_2	Nhiệt độ trung bình từ năm 1938-1977	10.218
s_1	Độ lệch chuẩn mẫu từ năm 1978-2018	10.433
s_2	Độ lệch chuẩn mẫu từ năm 1938-1977	10.713
n_1	Số phần tử của mẫu năm 1978-2018	492
n_2	Số phần tử của mẫu năm 1938-1977	477
T_0	Thống kê kiểm định	0.3656
v	Bậc tự do	963
$t_{0.05,\nu}$	Giá trị quyết định kiểm định	1.645

Ý nghĩa của các kết quả thu được:

Những kết quả μ_1, μ_2, s_1, s_2 thể hiện đúng bản chất của định nghĩa. n_1 và n_2 tương ứng là số tháng có dữ liệu từ năm 1938-1977 và 1978-2018.

Từ kết quả trên ta nhận thấy $T_0 = 0.3656 < t_{0.05,\nu} (= 1.645)$ do đó không đủ điều kiện để bác bỏ H_0, nghĩa là giả thuyết H_0 được chấp nhận. Hay: **Nhiệt độ trung bình từ năm 1978 đến năm 2018 nhỏ hơn nhiệt độ trung bình từ năm 1938 đến năm 1977 tại khu vực sân bay đảo Grand, bang Nebraska, Hoa Kỳ (Nghĩa là thời tiết có xu hướng trở lạnh từ năm 1978 đến 2018 tại khu vực này)**

References

[1] Cơ quan Quản lý Khí quyển và Đại dương Quốc gia (NOAA) - https://www.ncdc.noaa.gov/cdo-web/datasets

[2] Giáo trình Xác suất thống kê - Nguyễn Đăng Minh: https://drive.google.com/file/d/1DDewaUH1jb_u-81bluPl7MbSBy0C8j7B/view?usp=sharing