Suy luận và kiểm chứng giả thuyết về sự chênh lệch lượng mưa ở 2 bang California và Nevada bằng phương pháp kiểm định thống kê

Trịnh Quốc Huy - 20120013
Nguyễn Anh Khoa-20120118
Võ Thị Phước Thảo - 20120191

Ngày 10 tháng 4 năm 2022

1 Giới thiệu:

Lượng mưa - một đại lượng quen thuộc biển thị mức độ mưa nhiều hay mưa ít trong một năm, thường được đo đạt bằng thiết bị được gọi là vũ kế(Gồm một ống thủy tinh đứng được chia vạch, kết quả đo sẽ là số mm của mực nước trong ống). Và để xét lượng mưa trong một năm, người ta có nhiều cách để tiếp cận: Tính tổng lượng mưa của các ngày trong năm, tính tổng lượng mưa theo trung bình tháng, tính xem lượng mưa của ngày nhiều nhất trong năm, tính xem lượng mưa của tháng nhiều nhất trong năm,... Đại lượng mà bài toán này tập trung chính là lượng mưa cực đại trong 1 tháng trong năm, hiểu một cách đơn giản chính là tính tổng lượng mưa của các ngày trong tháng, kết quả là giá trị lớn nhất trong tất cả các giá trị của 12 tháng. Trong các giá trị khảo sát, có một vài năm chúng ta không thể có được kết quả vì nhiều lí do, nhưng vì xét trên thời gian dài là 100 năm, sẽ hạn chế được sai số trong việc đưa ra kết quả cuối.
Vấn đề đặt ra cho bài toán này chính là chứng minh rằng, khu vực cắm trại ranger station, California có lượng mưa tối đa trung bình lớn hơn khu vực Caliete thuộc bang Nevada trong khoảng thời gian thế kỉ 1900-2000. Đây là hai khu vực đại diện cho 2 miền của nước Mỹ. Đây là bài toán suy luận và để giải quyết vấn đề này, ta phải thực hiện kiểm định thống kê(Lấy dữ liệu từ năm 1900-2000) trên trường hợp so sánh 2 mẫu có $n \geq 30$ với độ lệch chuẩn δ chưa biết

2 Thông tin chung:

- Thời điểm: năm 1900-2000

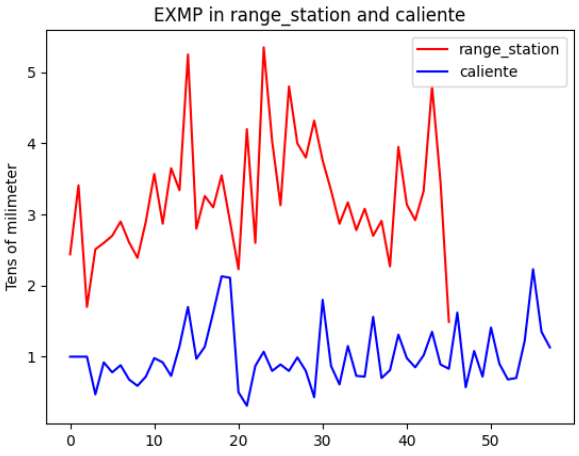
- Địa điểm: khu cắm trại ranger station, California và Caliete thuộc bang Nevada
- Dạng bài toán: Suy luận, tức đặt ra giả thuyết và kiểm chứng
- Thông tin kiểm định: EMXP(Extreme maximum precipitation) lượng mưa cực đại tháng tối đa(theo tổng các ngày trong tháng) trong năm 1900-2000, đơn vị tmm(tenths of milimeters)
- Phạm vi trong hai khu vực, xét trong toàn thế kỉ XX.

3 Dữ liệu:

Lấy dữ liệu từ trang web NCEI của Hoa Kỳ. Chọn mục "Global summary of the year" -> NCEI data search -> Chọn thông tin cần thiết để lấy dữ liệu(Mốc thời gian, thuộc tính quan tâm, năm bắt đầu, năm kết thúc, địa điểm,...) Dữ liệu được gửi tự động về mail

Tiền xử lý dữ liệu gồm có hai bước: Đọc dữ liệu từ file csv dưới sự trợ giúp của thư viện csv (Hàm đọc sẽ nhận về danh sách các list, ta chỉ giữ lại thông tin cần thiết cho thống kê). Dữ liệu sẽ bị khuyết ở một vài năm, cách tốt nhất chúng ta nên làm là bỏ qua

Các trường trong data là: Station: mã trạm, Name: tên trạm, Latitude, Longitude và elevation lần lượt là vĩ độ, kinh độ và chiều cao so với mực nước biển, date: Năm ghi nhận, EMXP(lượng mưa cực đại tháng tối đa(theo tổng các ngày trong tháng)), EMXP ATTRIBUTE: tháng mà lượng mưa đó xảy ra trong năm



Hình 1: Phân bố dữ liệu

4 Cơ sở lý thuyết:

Mô hình kiểm định thống kê được sử dụng trong trường hợp này là mô hình dùng để so sách hai mẫu khi chưa biết độ lệch chuẩn δ và có kích thước mẫu

$n1, n2 \geq 30$

Dựa vào giả thiết bài toán, ta có thể xác định giả thuyết H0 và đối thuyết H1 như sau:

$$\begin{cases} H0 : \overline{X1} = \overline{X2} \\ H1 : \overline{X1} > \overline{X2} \end{cases}$$

Xét kiểm định thống kê:

$$Z = \frac{\overline{X1} - \overline{X2}}{\sqrt{\frac{S1^2}{n1} + \frac{S2^2}{n2}}} \tag{1}$$

Xét với mức ý nghĩa alpha(nghĩa là khả năng kết quả quan sát sự khác biệt được nhìn thấy trên số liệu, ví dụ alpha=0.05 thì khả năng có sự khác biệt kết quả là 5%) nếu:

$$Z > Z_{1-alpha}$$

Trong đó $Z_{1-alpha}$ (1-alpha là giá trị mà hàm phân phối tích lũy của z tại đó nhận được)và nó cũng giá trị được trích từ bảng phân phối Gauss. Nếu biểu thức trên xảy ra, ta bác bỏ H0

5 Triển khai:

Ta chỉ cần dựa trên các công thức được nêu ở phần lý thuyết để kiểm chứng giả thuyết đưa ra

- Hàm means: Tính giá trị trung bình dưa trên mẫu dữ liệu. Đầu vào của hàm là một danh sách, biến S sẽ tính tổng giá trị trong danh sách sau đó chia cho chiều dài của danh sách
- Hàm S: Tính độ lệch chuẩn mẫu. Đầu vào của hàm cũng là một danh sách. Biến S tính tổng bình phương độ lệch của mỗi giá trị so với giá trị trung bình. Sau khi lấy giá trị này chia cho chiều dài của danh sách ta được phương sai mẫu, lấy căn bậc 2 ta được độ lệch chuẩn mẫu
- Hàm z: Tính giá trị thống kê kiểm định tương ứng của mẫu, ta thực hiện việc tính toán như công thức (1) ở phần trên. Các giá trị trung bình và độ lệch chuẩn mẫu được tính từ hàm means và hàm S.
- Hàm kiểm định: Ta tính giá trị thống kê kiểm định từ hàm z, sau đó tính giá trị $z_(1 - alpha)$ tương ứng và tiến hành so sánh, nếu giá trị thống kê kiểm định lớn hơn thì ta bác bỏ H0 và nếu không thì không đủ cơ sở bác bỏ H0. Trong hàm này có sử dụng hàm st.norm.ppf(1-alpha) từ thư viện scipy, nó giúp chúng ta lấy được giá trị z khi biết xác suất.

6 Kết quả

Mức ý nghĩa(alpha)	Z	$Z_{1\text{-alpha}}$	Kết luận
0.2	16.68493626820391	0.8416212335729143	Luong mua toi da trung binh cua khu vuc trong nam cua range station lon hon caliente
0.15	16.68493626820391	1.0364333894937898	Luong mua toi da trung binh cua khu vuc trong nam cua range station lon hon caliente
0.1	16.68493626820391	1.2815515655446004	Luong mua toi da trung binh cua khu vuc trong nam cua range station lon hon caliente
0.05	16.68493626820391	1.6448536269514722	Luong mua toi da trung binh cua khu vuc trong nam cua range station lon hon caliente
0.02	16.68493626820391	2.0537489106318225	Luong mua toi da trung binh cua khu vuc trong nam cua range station lon hon caliente
0.01	16.68493626820391	2.3263478740408408	Luong mua toi da trung binh cua khu vuc trong nam cua range station lon hon caliente
0.001	16.68493626820391	3.090232306167813	Luong mua toi da trung binh cua khu vuc trong nam cua range station lon hon caliente
0.0001	16.68493626820391	3.719016485455709	Luong mua toi da trung binh cua khu vuc trong nam cua range station lon hon caliente
0.00001	16.68493626820391	4.264890793923841	Luong mua toi da trung binh cua khu vuc trong nam cua range station lon hon caliente

Hình 2: Bảng kết quả

Với các mức ý nghĩa như trên, giá trị thống kê kiểm định cao hơn hẳn so với $_{1-alpha}$ ta có thể nói rằng, lượng mưa tối đa trung bình khu vực ranger camping(california) cao hơn so với Caliente(Nevada) trong thế kỉ 19. Có nghĩa là trung bình lượng mưa tháng cao nhất hằng năm của ranger camping cao hơn so với Caliente

7 Tài liệu tham khảo:

[1] Nguyễn Thị Mộng Ngọc, University of Science, VNU - HCM, slide Chương 5: Lý Thuyết Mẫu, Lý Thuyết ước lượng.

[2] Nguyễn Thị Mộng Ngọc, University of Science, VNU - HCM, slide Chương 6: Kiểm định giả thuyết thống kê