****

**REPORT**

**COMPUTER PROJECT – MAS291**

**XÁC SUẤT THỐNG KÊ**



**ĐỀ TÀI**

WEATHER AROUND SZEGED, HUNGARY IN APRIL 2006

Class: SE1609

Group: Nhóm 2

Lecturer: Dr.Nguyễn Việt Anh

Member:

* HE161905 - Nguyễn Sơn Tùng *(Leader)*
* HE151400 - Phạm Ngọc Sáng
* HE150498 - Nguyễn Quang Thịnh
* HE150933 - Nguyễn Thành Nam

**Mục Lục**

1. **NỘI DUNG**
2. **NGUỒN DỮ LIỆU**
   1. Thông tin về các nguồn dữ liệu
      1. Thông tin về nguồn dữ liệu
      2. Link nguồn dữ liệu
   2. Dữ liệu thu thập được
3. **THỐNG KÊ MÔ TẢ**
   1. Số liệu tính toán được
   2. Phân tích số liệu thống kê
4. **CÁC DẠNG TOÁN**
   1. Xây dựng khoảng tin cậy
   2. Kiểm định giả thiết
5. **KẾT LUẬN**
6. **NỘI DUNG**
   1. **Giới thiệu chung và Nội dung đề tài**

* Hungary là một quốc gia thuộc vùng Trung Âu, không giáp biển.
* Hungary có khí hậu ôn đới, chịu ảnh hưởng của khí hậu lục địa và Địa Trung Hải, vì vậy đất nước này sở hữu nền khí hậu ẩm ướt và được chia ra làm 4 mùa rõ rệt trong năm.
* Ở đây chúng ta sẽ tập trung phân tích vào mùa ẩm ướt (Tháng 4,5) của Hungary, đối tượng là thành phố Szeged.
* Đề tài sẽ thống kê về số liệu về Nhiệt độ (Temperature (C)), Nhiệt độ biểu kiến (Apparent Temperature(C)) và Tốc độ gió (Wind Speed (km/m)) trong khoảng thời gian Tháng 4 năm 2006 tại Szeged.
* Nhóm thu thập dữ liệu và thống kê sau đó nhóm sẽ làm sạch dữ liệu và tính các tham số đặc trưng để mô hình hóa bài toán thực tế thành bài toán có các dữ liệu đầy đủ và giải chúng theo các phương pháp của môn xác suất thống kê. Và đưa ra các giả thuyết, lời khẳng định, và sử dụng các phương pháp đã học để kiểm định các giả thuyết mà nhóm đã đề ra.
* Cuối cùng là đánh giá một lần tổng quan về dữ liệu mà nhóm đã thu thập được và sau khi làm sạch và phân tích dữ liệu.

1. **NGUỒN DỮ LIỆU**
   1. **Thông tin về nguồn dữ liệu**
      1. Kaggle Datasets

* Là một trang web trực tuyến tập trung vào việc sưu tầm các dữ liệu để giúp đỡ các học sinh, sinh viên có thể dễ dàng tiếp cận hơn các dataset lớn
* [Kaggle](https://www.kaggle.com/) được biết đến là một công ty con, hoạt động dưới sự điều hành, quản lý của tập đoàn Google LLC - một tập đoàn về công nghệ đa quốc gia được thành lập và có trụ sở chính tại Mỹ
* [Kaggle](https://www.kaggle.com/), được cập nhật bởi cộng đồng những người làm việc trong lĩnh vực ML, AI mỗi ngày, là một trong những thư viện tập dữ liệu trực tuyến lớn nhất.
* [Kaggle](https://www.kaggle.com/) là một nền tảng học máy dựa vào cộng đồng. website này chứa rất nhiều hướng dẫn bao gồm hàng trăm vấn đề machine learning trong cuộc sống thực khác nhau.

2.1.2 Link nguồn dữ liệu

[Weather in Szeged 2006-2016 | Kaggle](https://www.kaggle.com/datasets/budincsevity/szeged-weather)

* 1. **Dữ liệu thu thập được**
* Số liệu về tổng quan thời tiết, nhiệt độ, nhiệt độ biểu kiến, tốc độ gió của các ngày trong tháng 4 năm 2006 tại thành phố Szeged, Hungary được lấy tại Kaggle.

1. **THỐNG KÊ MÔ TẢ**
   1. **Số liệu tính toán được**

|  | Temperature (C) | Apparent Temperature (C) | Wind Speed (km/h) |
| --- | --- | --- | --- |
| Mean | 12.714405360134 | 12.2320351758794 | 10.7254721105528 |
| Standard Error | 0.205421908285273 | 0.239012647337818 | 0.355792843151858 |
| Median | 12.2666667 | 12.2666667 | 9.66 |
| Mode | 11.183333333 | 11.183333333 | 1.5295 |
| Standard Deviation | 4.0981541992241 | 4.76828733863617 | 7.09804492805989 |
| Sample Variance | 16.7948678406181 | 22.736564143798 | 50.3822418007568 |
| Kurtosis | -0.660511019386189 | -0.61630937392647 | -0.0920264997992364 |
| Skewness | 0.313592768660304 | -0.084082704450752 | 0.748653117564258 |
| Range | 19.1944444444444 | 21.4222222222222 | 32.1678 |
| Minimum | 3.72222222222222 | 1.49444444444444 | 0 |
| Maximum | 22.9166666666666 | 22.9166666666666 | 32.1678 |
| Sum | 5060.33333333332 | 4868.34999999998 | 4268.7379 |
| Count | 398 | 398 | 398 |
| Confidence Level (95.0%) | 0,040385073 | 0.4698886894118 | 0.699473582811831 |

*Các số liệu của Temperature, Apparent Temperature, Wind Speed tại Szeged*

* 1. **Phân tích số liệu thống kê**
* Mean
* Nhiệt độ trung bình (Temperature) với Nhiệt độ biểu kiến trung bình (Apparent Temperature) gần bằng nhau nhưng Nhiệt độ biểu kiến thấp hơn một chút xấp xỉ 0,5 so với với Nhiệt độ trung bình.
* Tốc độ gió đạt ngưỡng 10,7 km/h
* Standard Deviation (Độ lệch chuẩn)
* Là đại lượng thống kê dùng để đo mức độ phân tán và được tính bằng căn bậc hai của phương sai của dữ liệu.
* Độ lệch chuẩn về Temperature thấp hơn xấp xỉ khoảng 0,03 mặc dù Mean lại lớn hơn so với Apparent Temperature.
* Độ lệch chuẩn của Wind Speed theo tính toán được là xấp xỉ 7,1 km/h.
* Sample Variance (Phương Sai)
* Phương sai của một bảng số liệu là số đặc trưng cho độ phân tán của các số liệu so với số trung bình của nó.
* Phương sai của Temperature và Apparent Temperature cho số liệu bằng nhau. Đạt giá trị chính xác là: 12.2666666666666
* Mặc dù giá trị trung bình của Wind Speed chỉ là 10,7 những Phương sai đạt ngưỡng 50,3 km/h.
* Max, Min (Giá trị lớn nhất và thấp nhất)
* Giá trị Max về Temperature và Apparent Temperature đạt giá trị xấp xỉ bằng nhau là 22,9167 độ C.
* Trong khi đó thì giá trị Min giữa Temperature và Apparent Temperature có sự chênh lệch khá đáng kể: 2.2278 độ C.
* Wind Speed Max có thể lên tới 32,1678 km/h (Tốc độ gió cấp 5, có thể gây nguy hiểm). Và giá trị Min bằng 0 km/h vào hai ngày là:
  + 23/04/2006 lúc 21:00:00
  + 24/04/2006 lúc 06:00:00
* Và ngoài ra còn có khoảng 7 ngày WindSpeed chỉ đạt ngưỡng 0,3 - 0,6:
  + 11/04/2006 lúc 11:00:00 với WS = 0,644 km/h
  + 11/04/2006 lúc 12:00:00 với WS = 0,6762 km/h
  + 13/04/2006 lúc 15:00:00 với WS = 0,322 km/h
  + 20/04/2006 lúc 07:00:00 với WS = 0,3381 km/h
  + 21/04/2006 lúc 21:00:00 với WS = 0,6601 km/h
  + 22/04/2006 lúc 22:00:00 với WS = 0,5152 km/h
  + 23/04/2006 lúc 19:00:00 với WS = 0,322 km/h
  + 23/04/2006 lúc 20:00:00 với WS = 0,5313 km/h

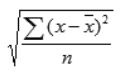
**4. CÁC DẠNG BÀI TOÁN**

|  |  | **SAMPLE DATA** |  |
| --- | --- | --- | --- |
|  | Temperature (C) | Apparent Temperature (C) | Wind Speed (km/h) |
| 1 | 9.472222222 | 7.388888889 | 14.1197 |
| 2 | 9.355555556 | 7.227777778 | 14.2646 |
| 3 | 9.377777778 | 9.377777778 | 3.9284 |
| 4 | 8.288888889 | 5.944444444 | 14.1036 |
| 5 | 8.755555556 | 6.977777778 | 11.0446 |
| 6 | 9.222222222 | 7.111111111 | 13.9587 |
| 7 | 7.733333333 | 5.522222222 | 12.3648 |
| 8 | 8.772222222 | 6.527777778 | 14.1519 |
| 9 | 10.82222222 | 10.82222222 | 11.3183 |
| 10 | 13.77222222 | 13.77222222 | 12.5258 |
| 11 | 16.01666667 | 16.01666667 | 17.5651 |
| 12 | 17.14444444 | 17.14444444 | 19.7869 |
| 13 | 17.8 | 17.8 | 21.9443 |
| 14 | 17.33333333 | 17.33333333 | 20.6885 |
| 15 | 18.87777778 | 18.87777778 | 15.3755 |
| 16 | 18.91111111 | 18.91111111 | 10.4006 |
| 17 | 15.38888889 | 15.38888889 | 14.4095 |
| 18 | 15.55 | 15.55 | 11.1573 |
| 19 | 14.25555556 | 14.25555556 | 8.5169 |
| 20 | 13.14444444 | 13.14444444 | 7.6314 |
| 21 | 11.55 | 11.55 | 7.3899 |
| 22 | 11.18333333 | 11.18333333 | 4.9266 |
| 23 | 10.11666667 | 10.11666667 | 6.6493 |
| 24 | 10.2 | 10.2 | 3.9284 |
| 25 | 10.42222222 | 10.42222222 | 16.9855 |
| 26 | 9.911111111 | 7.566666667 | 17.2109 |
| 27 | 11.18333333 | 11.18333333 | 10.8192 |
| 28 | 7.155555556 | 5.044444444 | 11.0768 |
| 29 | 6.111111111 | 4.816666667 | 6.6493 |
| 30 | 6.788888889 | 4.272222222 | 13.0088 |
| 31 | 7.261111111 | 5.155555556 | 11.1734 |
| 32 | 7.8 | 5.527777778 | 12.8156 |
| 33 | 9.872222222 | 7.933333333 | 13.7494 |
| 34 | 12.22222222 | 12.22222222 | 15.6331 |
| 35 | 15.09444444 | 15.09444444 | 17.549 |
| 36 | 17.35555556 | 17.35555556 | 22.7815 |
| 37 | 19.00555556 | 19.00555556 | 23.8924 |
| 38 | 20.04444444 | 20.04444444 | 28.3682 |
| 39 | 21.05 | 21.05 | 26.9031 |
| 40 | 21.18333333 | 21.18333333 | 25.6956 |
| 41 | 20.11666667 | 20.11666667 | 25.3092 |
| 42 | 20.21666667 | 20.21666667 | 18.1125 |
| 43 | 20 | 20 | 23.4255 |
| 44 | 17.8 | 17.8 | 20.0445 |
| 45 | 16.06111111 | 16.06111111 | 21.3969 |
| 46 | 15.02222222 | 15.02222222 | 21.3808 |
| 47 | 14.42222222 | 14.42222222 | 20.0123 |
| 48 | 14.25555556 | 14.25555556 | 16.5025 |
| 49 | 13.77222222 | 13.77222222 | 17.0982 |
| 50 | 13.28333333 | 13.28333333 | 14.3612 |
| 51 | 8.633333333 | 5.466666667 | 22.0409 |

*Số liệu thống kê lượng mưa, lượng mưa biểu kiến và tốc độ gió*

**4.1 Xây dựng khoảng tin cậy giá trị trung bình của tổng thể**

# **4.1.1** Xây dựng các khoảng tin cậy hai phía 95% cho chênh lệch tỷ lệ nhiệt độ (Temperature).

* Ta có:
* khoảng tin cậy = 95% ⇒ 𝛂 = 0.05
* = 13.31546841
* s =  = 4.414932338

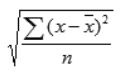
⇒ E =Ta/2 \* = 1.241719729

⇒ Khoảng tin cậy (12.07374868;14.55718814)

Kết luận: Chúng tôi tin rằng 95% rằng nhiệt độ trung bình nằm trong khoảng từ 12.07374868 đến 14.55718814 độ C

**4.1.2** Xây dựng các khoảng tin cậy hai phía 95% cho chênh lệch tỷ lệ nhiệt độ biểu kiến (Apparent temperature).

Ta có:

* khoảng tin cậy = 95% ⇒ 𝛂 = 0.05
* =12.67527233
* s =  = 5.194191118

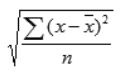
⇒ E = Ta/2 \* = 1.475426368

⇒ Khoảng tin cậy (11.21438254; 14.13616212)

Kết luận: Chúng tôi tin rằng 95% rằng nhiệt độ trung bình biểu kiến nằm trong khoảng từ 11.21438254 đến 14.13616212 độ C

**4.1.3** Xây dựng các khoảng tin cậy hai phía 95% cho chênh lệch tốc độ gió (Wind speed (km/h)).

Ta có:

* khoảng tin cậy = 95% ⇒ 𝛂 = 0.05
* = 15.4146451
* s =  = 6.079492393

⇒ E = Ta/2 \* = 1.709884789

⇒ Khoảng tin cậy (13.70476031;17.12452989)

Kết luận: Chúng tôi tin rằng 95% rằng tốc độ gió nằm trong khoảng từ 11.19984596 đến 14.1506987 độ C

**4.2 Kiểm định giả thuyết**

|  | ***Temperature ©*** | ***Apparent Temperature ©*** |
| --- | --- | --- |
|  |
| Mean | 12.71440536 | 12.23203518 |
| Standard Error | 0.2054219083 | 0.2390126473 |
| Median | 12.26666667 | 12.26666667 |
| Mode | 11.18333333 | 11.18333333 |
| Standard Deviation | 4.098154199 | 4.768287339 |
| Sample Variance | 16.79486784 | 22.73656414 |
| Kurtosis | -0.6605110194 | -0.6163093739 |
| Skewness | 0.3135927687 | -0.08408270445 |
| Range | 19.19444444 | 21.42222222 |
| Minimum | 3.722222222 | 1.494444444 |
| Maximum | 22.91666667 | 22.91666667 |
| Sum | 5060.333333 | 4868.35 |
| Count | 398 | 398 |
| Confidence Level(95.0%) | 0.4038507265 | 0.4698886894 |

# **4.2.1**

# **Bài toán:** Có bằng chứng nào cho thấy rằng nhiệt độ thực tế (Temperature) chênh lệch so với nhiệt độ biểu kiến (Apparent temperature) là 14 độ hay không?

|  | ***Temperature*** | ***Apparent Temperature*** |
| --- | --- | --- |
| **Mean** | 397,249032 | 383,9556235 |
| **Variance** | 1812132,814 | 1676713,897 |
| **Observations** | 14 | 14 |
| **Hypothesised Mean Difference** | 14 |  |
| **df** | 26 |  |
| **t Stat** | -0,00141544003 |  |
| **P(T<=t) one-tail** | 0,4994407236 |  |
| **t Critical one-tail** | 1,70561792 |  |
| **P(T<=t) two-tail** | 0,9988814472 |  |
| **t Critical two-tail** | 2,055529439 |  |

H0 = **µ**1 **- µ**2  **= 14**

H1 = **µ**1 **- µ**2  **≠ 14**

**To =** -0.00141544003

**T**a/2 **=** 2.055529439

⇒ Không thể bác bỏ H0

Từ kết quả trên, chúng tôi tin rằng nhiệt độ thực tế (Temperature) chênh lệch so với nhiệt độ biểu kiến (Apparent temperature) là 14 độ

# **4.2.2**

# **Bài toán:** Có bằng chứng nào cho thấy rằng nhiệt độ thực tế (Temperature) bằng 20 độ C hay không?

*= 0.05*

Ta có:

Test claim: = 0.05, = 16.79486784,

: = 20

: 20

= 1.648681534

= ( - )/() = 19.91366006

⇒ Bác bỏ

Từ kết quả trên, chúng tôi tin rằng nhiệt độ thực tế khác 20 độ C

**4.3 Hồi quy tuyến tính**

| SUMMARY OUTPUT |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |
| *Regression Statistics* | |  |  |  |  |  |
| Multiple R | 0.99493 |  |  |  |  |  |
| R Square | 0.989886 |  |  |  |  |  |
| Adjusted R Square | 0.989465 |  |  |  |  |  |
| Standard Error | 0.453151 |  |  |  |  |  |
| Observations | 51 |  |  |  |  |  |
|  |  |  |  |  |  |  |
| ANOVA |  |  |  |  |  |  |
|  | *df* | *SS* | *MS* | *F* | *Significance F* |  |
| Regression | 2 | 964.7247908 | 482.3624 | 2349.028 | 1.31E-48 |  |
| Residual | 48 | 9.856586675 | 0.205346 |  |  |  |
| Total | 50 | 974.5813774 |  |  |  |  |
|  |  |  |  |  |  |  |
|  | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
| Intercept | 2.084899 | 0.189349033 | 11.01088 | 9.86E-15 | 1.704186958 | 2.46561 |
| Apparent Temperature (C) | 0.78148 | 0.01541882 | 50.6835 | 2.43E-43 | 0.750478097 | 0.812481 |
| Wind Speed (km/h) | 0.085964 | 0.013173518 | 6.525501 | 3.93E-08 | 0.05947667 | 0.112451 |

chỉ ra mối quan hệ tuyến tính mạnh mẽ giữa hai biến này vì mô hình hồi quy đã giải thích 98,9886% độ biến thiên trong dự đoán nhiệt độ (chỉ 1,0114% độ biến thiên của mẫu về nhiệt độ là do các yếu tố không được mô hình hồi quy xem xét)

The regression quotation is

= 2.08489 - + 0.085964

### ii. Interpret the regression coefficient (slope) in the context of your research.

Nếu Nhiệt độ biểu kiến tăng 10 độ C dẫn đến nhiệt độ trung bình dự đoán sẽ giảm khoảng 7.814 độ C, với tốc độ gió không đổi

Nếu Tốc độ gió tăng 10 km / h, dự đoán nhiệt độ trung bình sẽ tăng 0,8 độ C,với nhiệt độ biểu kiến không đổi

**5. KẾT LUẬN**

Đề tài đã đạt được các mục tiêu đề ra trong quá trình nghiên cứu và phân tích thông qua các vấn đề đặt ra. Chúng tôi đã phân tích được mối tương quan giữa các biến, xây dựng khoảng tin cậy cho giá trị trung bình của tổng thể. Chúng tôi cũng đã nghiên cứu các thông tin về nguồn dữ liệu và thống kê mô tả dữ liệu thu thập được.

Bên cạnh đó là các bài toán kiểm định giả thuyết của Nhiệt độ thực tế (Temperature) và nhiệt độ biểu kiến (Apparent temperature), dựa vào tính toán và biểu đồ Line Fit Plot để chỉ ra mối tương quan mạnh yếu giữa các biến và viết phương trình hồi quy tuyến tính, dựa vào phương trình tuyến tính để giải thích hệ số hồi quy trong bối cảnh của bài nghiên cứu.