**2.2:Đo lường xu hướng trung tâm: Trung bình, Trung vị và Chế độ(Mean, Median and Mode)**

**2.2.1:**

Trong phần này, chúng tôi xem xét các cách khác nhau để đo lường xu hướng trọng tâm của dữ liệu. Giả sử rằng chúng ta có một số thuộc tính X, như tiền lương, đã được ghi lại cho một tập hợp các đối tượng.

Gọi x1, x2, ..., xN là tập N giá trị quan sát được hoặc quan sát được đối với X. Ở đây, các giá trị này cũng có thể được gọi là tập dữ liệu (đối với X). Nếu chúng ta lập kế hoạch cho các quan sát đối với tiền lương, hầu hết các giá trị sẽ giảm ở đâu? Điều này cho chúng ta một ý tưởng về thời gian mười trung tâm của dữ liệu. Các thước đo về xu hướng trung tâm bao gồm giá trị trung bình, trung vị, chế độ và

tầm trung. Thước đo số phổ biến và hiệu quả nhất về "trung tâm" của một tập hợp dữ liệu là trung bình (số học). Gọi x1, x2, ..., xN là tập N giá trị hoặc quan sát, chẳng hạn như một số thuộc tính số X, như tiền lương. Giá trị trung bình của bộ giá trị này là: = (2.1)

Điều này tương ứng với hàm tổng hợp được tích hợp sẵn, trung bình (avg () trong SQL), được cung cấp trong các hệ thống cơ sở dữ liệu quan hệ.

**\*Ví Dụ 2.6:**

**\*Trung bình:**

Giả sử chúng ta có các giá trị sau cho tiền lương (tính bằng hàng nghìn đô la), được hiển thịtheo thứ tự tăng dần: 30, 36, 47, 50, 52, 52, 56, 60, 63, 70, 70, 110. Sử dụng Eq. (2.1), chúng tôi có:

===58

Do đó mức lương trung bình là: $58.000

Đôi khi, mỗi giá trị xi trong một tập hợp có thể được liên kết với trọng lượng wi

cho i = 1, ..., N.Các trọng số phản ánh ý nghĩa, tầm quan trọng hoặc tần suất xuất hiện gắn liền với giá trị tương ứng của chúng. Trong trường hợp này, chúng ta có thể tính:

= (2.2)

Đây được gọi là trung bình cộng có trọng số hoặc trung bình có trọng số.

Mặc dù giá trị trung bình là đại lượng hữu ích duy nhất để mô tả một tập dữ liệu, nhưng nó không luôn là cách tốt nhất để đo trung tâm của dữ liệu. Một vấn đề lớn với phương tiện là độ nhạy của nó đối với các giá trị cực đoan (ví dụ: ngoại lệ). Ngay cả một số nhỏ các giá trị cực đoan có thể làm hỏng giá trị trung bình. Ví dụ, mức lương trung bình tại một công ty về cơ bản có thể là được đẩy lên bởi một vài nhà quản lý được trả lương cao. Tương tự, điểm trung bình của một lớp trong

một kỳ thi có thể bị kéo xuống khá nhiều bởi một số điểm rất thấp. Để bù đắp hiệu ứng gây ra bởi một số lượng nhỏ các giá trị cực trị, thay vào đó, chúng tôi có thể sử dụng giá trị trung bình được cắt bớt, là giá trị trung bình thu được sau khi cắt bỏ các giá trị ở cực cao và cực thấp. Vì ví dụ, chúng tôi có thể sắp xếp các giá trị được quan sát cho tiền lương và loại bỏ 2% trên cùng và dưới cùng trước khi tính toán giá trị trung bình. Chúng ta nên tránh cắt bỏ một phần quá lớn (chẳng hạn như

20%) ở cả hai đầu, vì điều này có thể dẫn đến mất thông tin có giá trị. Đối với dữ liệu lệch (không đối xứng), thước đo trung tâm dữ liệu tốt hơn là giá trị trung bình,

là giá trị giữa trong một tập hợp các giá trị dữ liệu có thứ tự. Nó là giá trị phân tách

nửa cao hơn của tập dữ liệu từ nửa dưới.Trong xác suất và thống kê, trung vị thường áp dụng cho dữ liệu số; Tuy vậy, chúng tôi có thể mở rộng khái niệm sang dữ liệu thứ tự. Giả sử rằng một tập dữ liệu nhất định gồm N giá trị cho một thuộc tính X được sắp xếp theo thứ tự tăng dần. Nếu N là số lẻ, thì trung vị là giá trị giữa của tập hợp có thứ tự. Nếu N chẵn thì trung vị không phải là duy nhất; nó là hai giá trị ở giữa và bất kỳ giá trị nào ở giữa. Nếu X là một thuộc tính số trong trường hợp này, bằng cách quy ước, giá trị trung vị được lấy là giá trị trung bình của hai giá trị ở giữa.

**Ví dụ 2.7:**

**\*Trung Vị:**

Hãy tìm giá trị trung bình của dữ liệu từ Ví dụ 2.6. Dữ liệu đã được sắp xếp

theo thứ tự tăng dần. Có một số lượng quan sát chẵn (tức là, 12); do đó,

trung vị không phải là duy nhất. Nó có thể là bất kỳ giá trị nào trong hai giá trị ở giữa là 52 và 56 (nghĩa là trong giá trị thứ sáu và thứ bảy trong danh sách). Theo quy ước, chúng tôi chỉ định giá trị trung bình của hai giá trị ở giữa làm giá trị trung vị; đó là =54. Do vậy trung vị là $54.000.

Giả sử rằng chúng ta chỉ có 11 giá trị đầu tiên trong danh sách. Cho một số lẻ

giá trị, trung vị là giá trị ở giữa. Đây là giá trị thứ sáu trong danh sách này, có

trị giá $52.000 .

Trung vị rất tốn kém để tính toán khi chúng ta có một số lượng lớn các quan sát.

Tuy nhiên, đối với các thuộc tính số, chúng ta có thể dễ dàng ước tính giá trị. Giả sử rằng dữ liệu được nhóm trong các khoảng thời gian theo giá trị dữ liệu xi của chúng và tần suất (tức là số lượng giá trị dữ liệu) của mỗi khoảng đã biết. Ví dụ, nhân viên có thể được nhóm theo mức lương hàng năm của họ trong các khoảng thời gian như $ 10–20,000, $ 20–30,000, và như thế. Gọi khoảng chứa tần số trung vị là giá trị trung vị. Chúng tôi có thể ước tính giá trị trung bình của toàn bộ tập dữ liệu (ví dụ: mức lương trung bình) bằng cách nội suy sử dụng công thức

Trung vi=L1+() \* width, (2.3)

trong đó L1 là biên dưới của khoảng trung bình, N là số giá trị trong toàn bộ tập dữ liệu, là tổng các tần số của tất cả các khoảng thấp hơn khoảng thời gian trung bình , (freq)median là tần số của khoảng thời gian trung bình, và width là chiều rộng của khoảng giữa.

Chế độ là một thước đo khác của xu hướng trung tâm. Chế độ cho một tập hợp dữ liệu là giá trị xuất hiện thường xuyên nhất trong tập hợp. Do đó, nó có thể được xác định cho các thuộc tính định lượng và chất lượng. Tần số lớn nhất có thể tương ứng với một số giá trị khác nhau, dẫn đến nhiều hơn một chế độ. Tập dữ liệu có một, hai, hoặc ba chế độ tương ứng được gọi là đơn phương thức, hai phương thức và ba phương thức. Nói chung, một tập dữ liệu có hai hoặc nhiều chế độ là đa phương thức. Ở một thái cực khác, nếu mỗi giá trị dữ liệu chỉ xảy ra một lần, sau đó không có chế độ.

**\*Ví dụ 2.8:**

**\*Mode:**

Dữ liệu từ Ví dụ 2.6 là hai phương thức. Hai chế độ là $ 52.000 và $70.000 .

Đối với dữ liệu số đơn phương thức bị lệch vừa phải (không đối xứng), chúng tôi có mối quan hệ thực nghiệm sau:

trung bình - chế độ ≈ 3 × (trung bình - trung vị). (2.4)

Điều này ngụ ý rằng chế độ cho các đường cong tần số đơn phương bị lệch vừa phải có thể dễ dàng gần đúng nếu biết giá trị trung bình và giá trị trung vị.

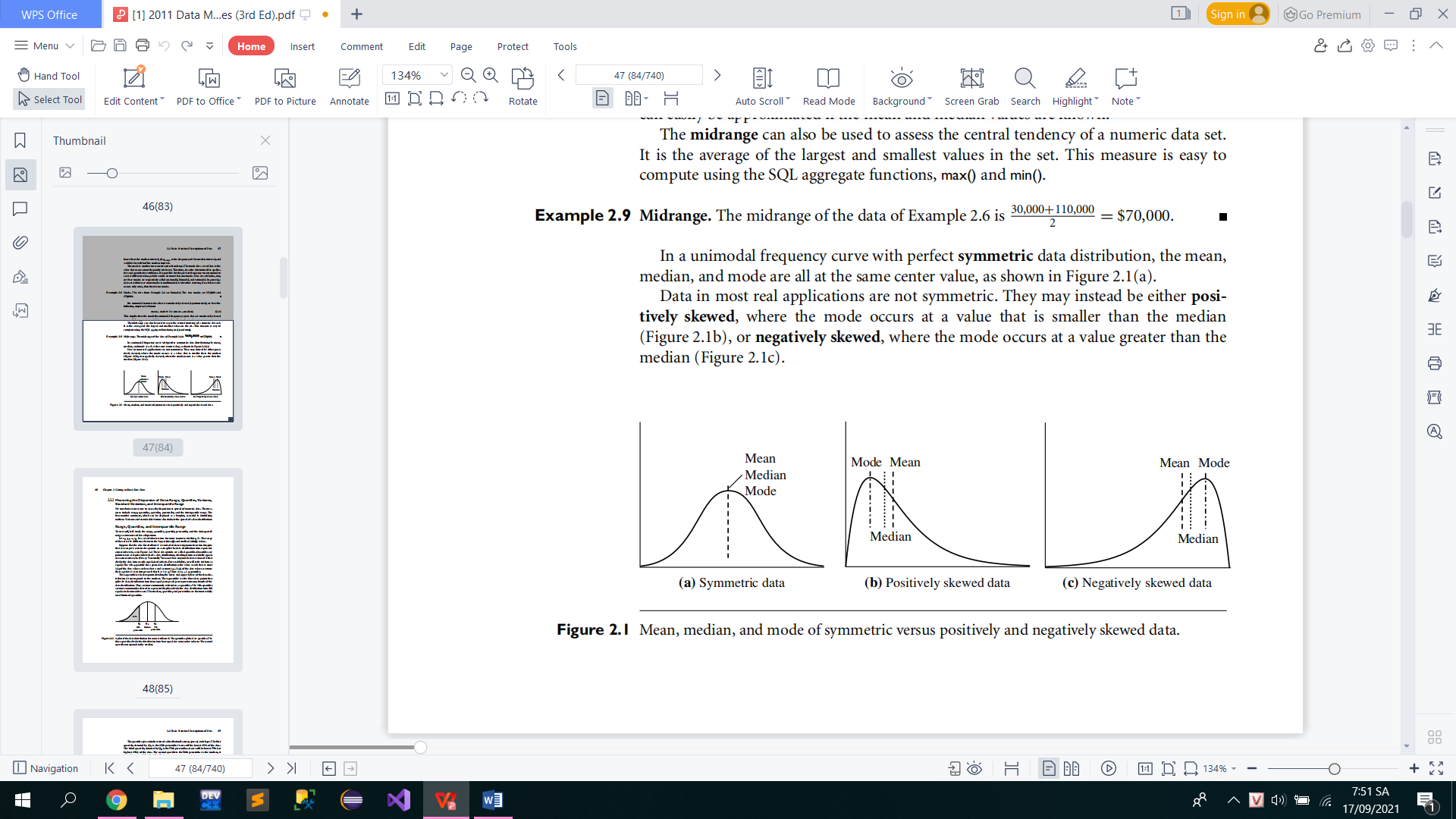
Khoảng giữa cũng có thể được sử dụng để đánh giá xu hướng trung tâm của tập dữ liệu số. Nó là giá trị trung bình của các giá trị lớn nhất và nhỏ nhất trong tập hợp. Biện pháp này rất dễ tính toán bằng cách sử dụng các hàm tổng hợp SQL, max () và min ().

**\*Ví dụ 2.9:**

**\*Tầm trung:**

Khoảng giữa của dữ liệu của Ví dụ 2.6 là

Trong đường cong tần số đơn phương thức với phân phối dữ liệu đối xứng hoàn hảo, giá trị trung bình trung vị và chế độ đều ở cùng một giá trị trung tâm, như thể hiện trong Hình 2.1 (a).Dữ liệu trong hầu hết các ứng dụng thực tế không đối xứng. Thay vào đó, chúng có thể bị lệch về phía trước, trong đó chế độ xảy ra ở giá trị nhỏ hơn giá trị trung bình (Hình 2.1b), hoặc lệch âm, trong đó chế độ xảy ra ở giá trị lớn hơn trung vị (Hình 2.1c)



**PHẦN BÀI TẬP**

**Bài tập 2.2:** Giả sử rằng dữ liệu để phân tích bao gồm tuổi thuộc tính. Giá trị độ tuổi cho dữ liệu các bộ giá trị là (theo thứ tự tăng dần) 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70.

**a)Tìm giá trị trung bình của dữ liệu và tìm giá trị trung vị của dữ liệu**

\*Giá trị trung bình của dữ liệu là:

\*Giá trị trung vị của dữ liệu là:

Vì số lượng mẫu quan sát là 27 (là số lẻ) nên giá trị trung vị sẽ là giá trị thứ 13 ứng với đó là số 25.

Vậy giá trị trung vị của dữ liệu là 25.

**b)Tìm mode của dữ liệu và nhận xét về phương thức của dữ liệu.**

\*Mode(Chế độ) của dữ liệu là:

trung bình - chế độ ≈ 3 × (trung bình - trung vị)

= 29.962 – chế độ ≈ 3 x (29.962-25)

* Chế độ ≈ 29.962-3x(29.962-25)
* Chế độ ≈15.076

\*Phương thức của dữ liệu là 2 phương thức.

**c)Tìm khoảng giữa của dữ liệu( Ta lấy giá trị đầu cộng giá trị cuối rồi chia cho 2)**

\*Khoảng giữa của dữ liệu là:

**d) Bạn có thể tìm (gần đúng) phần tư đầu tiên (Q1) và phần tư thứ ba (Q3) của dữ liệu không?**

**-**Ta sẽ lấy ra 25% dữ liệu thấp nhất: 13, 15, 16, 16, 19, 20, 20 (25% của 27 xấp xỉ 7). Sau đó ta sẽ lấy giá trị cao nhất của 25% dữ liệu này là 20

Suy ra Q1=20.

-**-**Ta sẽ lấy ra 75% dữ liệu thấp nhất: 13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35. Sau đó ta sẽ lấy giá trị cao nhất của 75% dữ liệu này là 35

Suy ra Q3=35.

**e) Đưa ra bản tóm tắt năm số của dữ liệu.(min,max,Q1,Q3)**

**-**Min=13.

-Max=70.

-Trung vị =25.

-Q1=20.

-Q3=35.

**Bài tập 2.3:** Giả sử rằng các giá trị của một tập dữ liệu nhất định được nhóm thành các khoảng. Khoảng thời gian và các tần số tương ứng như sau:

|  |  |
| --- | --- |
| Tuổi | Tần số |
| 1-5 | 200 |
| 6-15 | 450 |
| 16-20 | 300 |
| 21-50 | 1500 |
| 51-80 | 700 |
| 81-110 | 44 |

Tính giá trị trung bình gần đúng cho dữ liệu:

\*Giá trị trung bình gần đúng của dữ liệu là :

≈ 35.7

**Bài Tập 2.4:** Giả sử rằng một bệnh viện đã kiểm tra dữ liệu tuổi và lượng mỡ cơ thể cho 18 người trưởng thành được chọn ngẫu nhiên với các kết quả sau:

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | 23 | 23 | 27 | 27 | 39 | 41 | 47 | 49 | 50 |
| %fat | 9.5 | 26.5 | 7.8 | 17.8 | 31.4 | 25.9 | 27.4 | 27.2 | 31.2 |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| age | 52 | 54 | 54 | 56 | 57 | 58 | 58 | 60 | 61 |
| %fat | 34.6 | 42.5 | 28.8 | 33.4 | 30.2 | 34.1 | 32.9 | 41.2 | 35.7 |

a)Tính giá trị trung bình và độ lệch chuẩn của age và %fat

\*Giá trị trung bình:

\*Age:

≈46.444

\*%fat:

≈28.793

\*Độ lệch chuẩn:

\*Age:

Ta có:

(23-46.444)^2=549.621

(27-46.444)^2=378.069

(39-46.444)^2=55.413

(41-46.444)^2=29.637

(47-46.444)^2=0.309

(49-46.444)^2=6.533

(50-46.444)^2=12.645

(52-46.444)^2=30.869

(54-46.444)^2=57.093

(56-46.444)^2=91.317

(57-46.444)^2=111.429

(58-46.444)^2=133.541

(60-46.444)^2=183.765

(61-46.444)^2=211.877

Suy Ra:549.621\*2+378.069\*2+55.413+29.637+0.309+6.533+12.645+30.869+

57.093\*2+91.317+111.429+133.541\*2+183.765+211.877

=2970.442

* Phương sai là 2970.442 : (18-1) = 174.732
* Độ lệch chuẩn là : ≈13.218

\*%fat:

Ta có:

(9.5-28.793)^2=372.219

(26.5-28.793)^2=5.257

(7.8-28.793)^2=440.706

(17.8-28.793)^2=120.846

(31.4-28.793)^2=6.796

(25.9-28.793)^2=8.369

(27.4-28.793)^2=1.94

(27.2-28.793)^2=2.537

(31.2-28.793)^2=5.793

(34.6-28.793)^2=33.721

(42.5-28.793)^2=187.881

(28.8-28.793)^2=4.9\*10^(1:5)

(33.4-28.793)^2=21.224

(30.2-28.793)^2=1.979

(34.1-28.793)^2=28.164

(32.9-28.793)^2=16.867

(41.2-28.793)^2=153.933

(35.7-28.793)^2=47.706

Suy Ra:372.219+5.257+440.706+120.846+6.796+8.369+1.94+2.537+5.793+33.721+187.881+4.9\*10^(1:5)+21.224+1.979+28.164+16.867+153.933+47.706

=1452.938

* Phương sai là : 1452.938 : (18-1) ≈85.466
* Độ lệch chuẩn là : 9.244

**\*Tính trung vị**

**\*Age:**

Vì số lượng mẫu quan sát là 18 (là số chẵn) nên giá trị trung vị sẽ bằng trung bình cộng của hai giá trị ở giữa.

Trung vị cần tìm là:

**\*%fat:**

Vì số lượng mẫu quan sát là 18 (là số chẵn) nên giá trị trung vị sẽ bằng trung bình cộng của hai giá trị ở giữa.

Trung vị cần tìm là: