1

Tính trung bình (mean) : cộng lại chia n

Tính trung vị (median): n/2 (nếu 13.5 làm tròn lên là phần tử thứ 14)

Khoảng giữa (mid point) : (min + max)/2

Các loại mode : single modal, bimodal, triple modal, multi modal,…

Tính tứ phân vị :

Text, letter

Description automatically generated

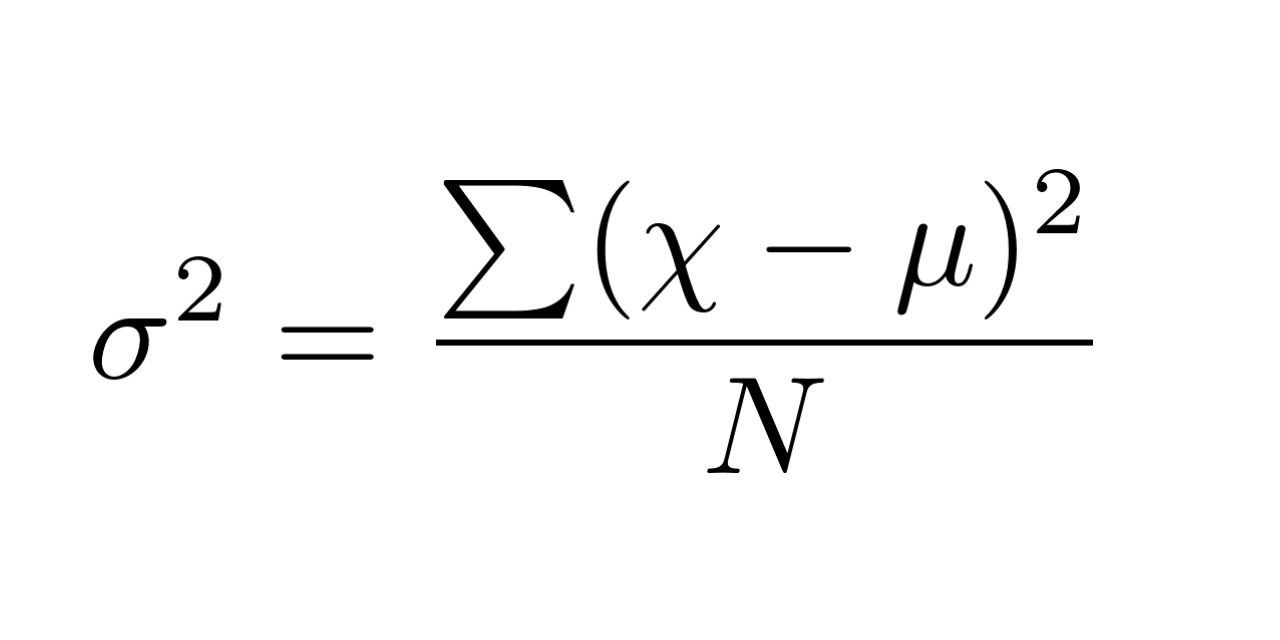
Boxplot:

Chart

Description automatically generated

Tính trung tâm : mean mode median

Tính phân bố : độ lệch chuẩn , phương sai

Phương sai : 

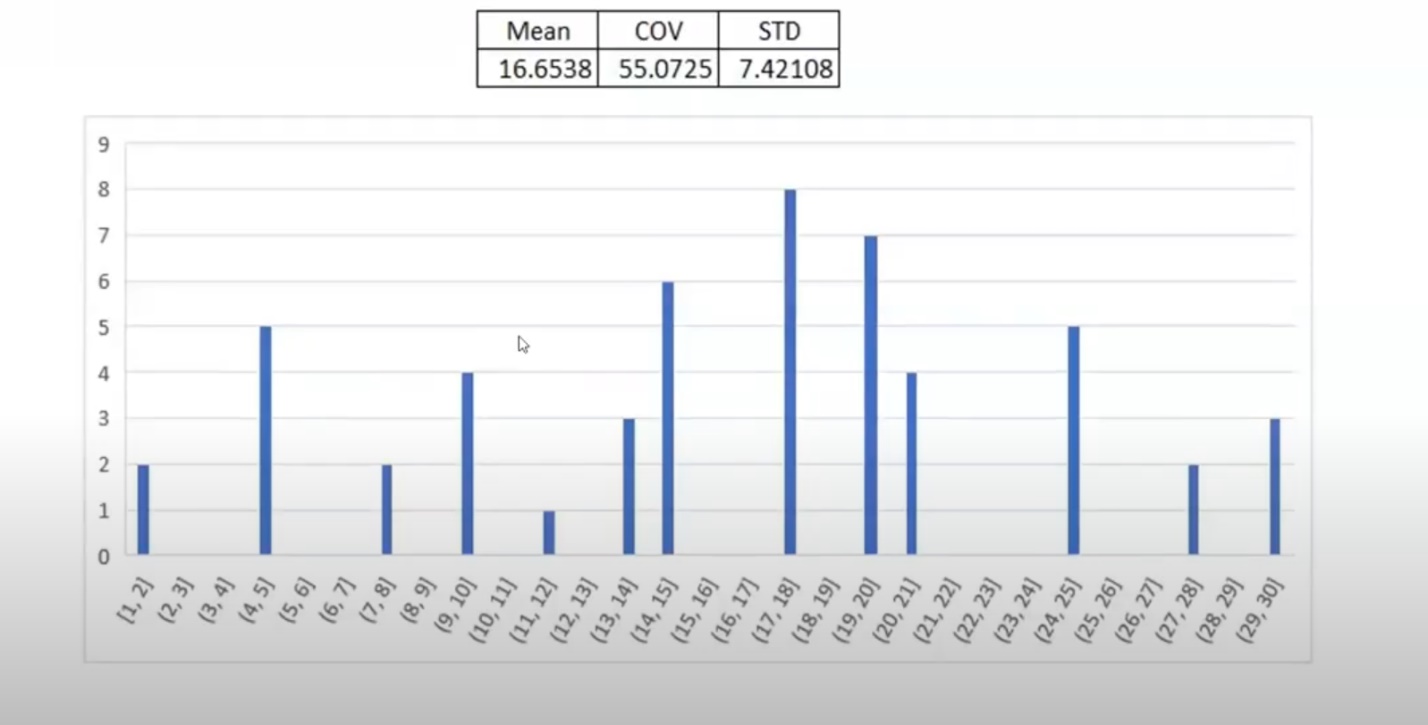
Độ lệch chuẩn :



Histogram : #bin-width=5 -> (30-1)/x=5 -> bins=6(đưa vào code để vẽ)

Chart, histogram

Description automatically generated



Chia giỏ theo độ rộng : max-min=độ rộng vd chia 3 giỏ

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.

(70-13)/3= 19

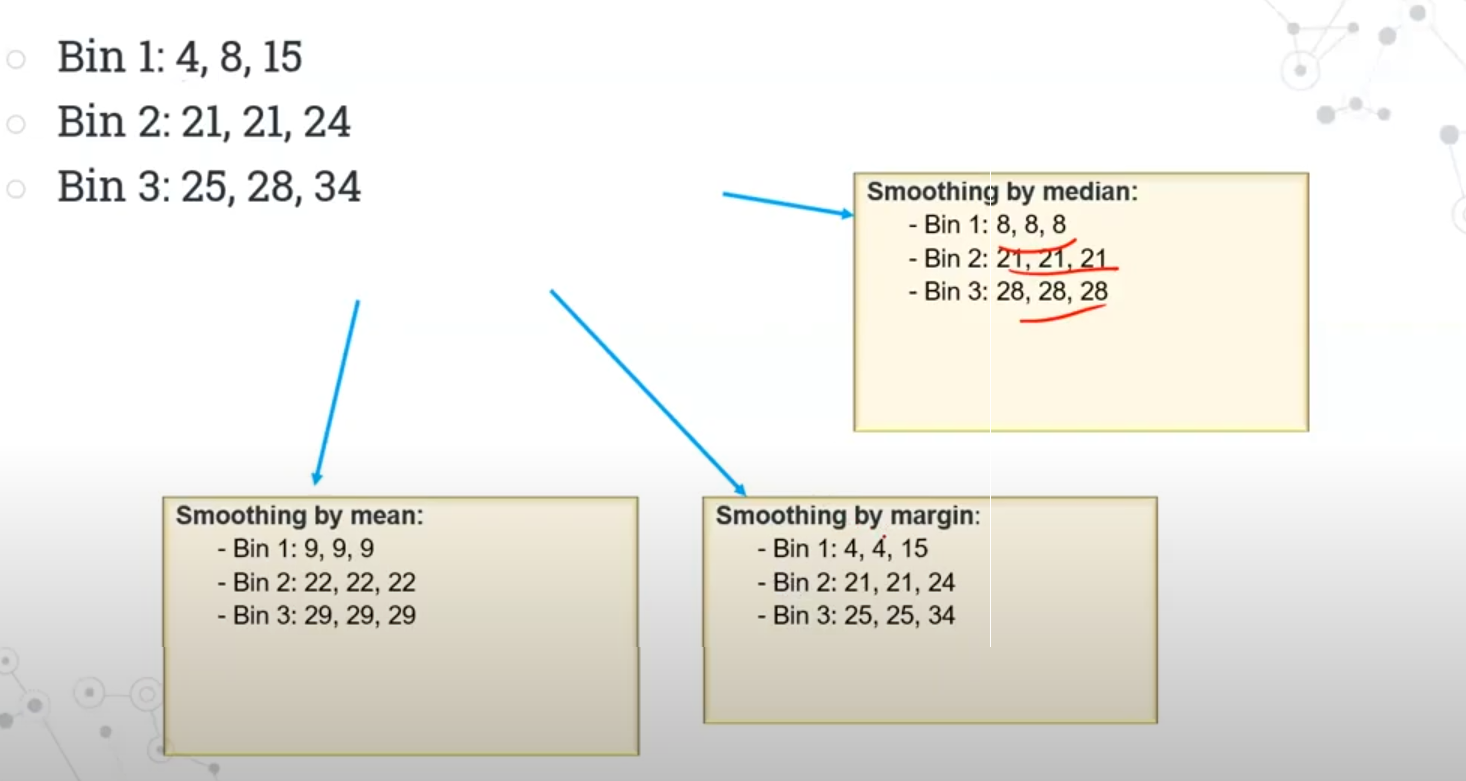
* Phạm vi=19 -> Diagram

  Description automatically generated
* 13+19=32 nếu thằng nào = 32 đưa vào giỏ tiếp theo….

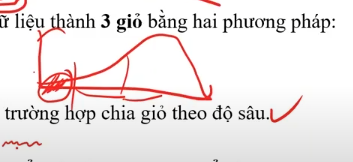
Chia giỏ theo độ sau : (n/3)=27/3 = mỗi giỏ 9 số, nếu chia ra lẻ thì nẳm ở giỏ cuối cùng

Làm trơn theo giá trị trung bình : lấy số phần tử cộng lại chia cho tổng số phần tử của mỗi giỏ

Vd : ở chia giỏ theo độ sâu lấy các phần tử mỗi giỏ cộng lại chia cho 9, tương tự giỏ 2 ,3



Phát hiện outlier : 3-sigma nếu phân phối chuẩn (mean trùng với median ) thì tính sigma những thằng nào nằm ngoài sigma là outlier

Nếu kkhong  xem giá trị mean là bao nhiều, nếu khoảng cách từ đuôi xa giá trị mean và median ở một khoảng cách nhất định nào đó thì cắt bỏ outlier này (hoặc thằng nào nằm ngoài q1 -> q3 -> outlier)

Chart, scatter chart

Description automatically generated

chuẩn hóa min-max= A picture containing text, device, gauge, meter

Description automatically generated

vd 200, 300, 400, 600, 1000

chuẩn hóa min-max -> 0, 0.125…

chuẩn hóa z-score: Diagram

Description automatically generated with low confidence

tính trung bình và độ lệch chuẩn(chia cho n) -> -1.07,…

chuẩn hóa điểm z bằng độ lệch tuyệt đối trung bình :

Text

Description automatically generated with medium confidence

Graphical user interface, application, table

Description automatically generated

A picture containing text, device, gauge

Description automatically generated

Thống nhất chia cho n, nếu trong mẫu có nói tỉ lệ này ảnh hưởng tới mẫu thì mới điều chỉnh n-1

Fill missing value : Nếu như là phân phối chuẩn thì điền mean sẽ ổn (phân phối có 1 mode)

Có thể điều chỉnh mean theo cái miền giá trị để vô đúng cái miền của nó (đa phân phối , xem thử thiếu nằm ở phân phối nào để tính mean theo phân phối đó )

Một số cách để làm cho dữ liệu có thể xử lý nhanh hơn : số hóa các name thành 0 1 2, no biến thành 0 yes thành 1 (binary) hoặc đổi đơn vi vd : 35000 -> 35 (vd tên quốc gia..)

Để đưa vào các hàm tính khoảng cách thì cần phải chuẩn hóa dữ liệu (z, min-max) vd : tuổi 35 salary tính theo ngàn thì khi đưa vào tính toán cái tuổi sẽ k ảnh hưởng gì nhiều đến việc tính toán nên cần phải chuẩn hóa (z-score, min-max…)để bảo đảm tuổi và salary ảnh hưởng như nhau

mean,mode:<https://www.calculatorsoup.com/calculators/statistics/mean-median-mode.php>

phương sai: <https://www.calculatorsoup.com/calculators/statistics/variance-calculator.php>

thống kê:<https://www.calculatorsoup.com/calculators/statistics/descriptivestatistics.php>

chuẩn hóa: z-score https://www.calculatorsoup.com/calculators/statistics/z-score-calculator.php