MÔN HỌC

THỐNG KÊ ỨNG DỤNG - XD (KC107)

GIÁO VIÊN GIẢNG DẠY

ĐẶNG THẾ GIA

Bộ môn Kỹ Thuật Xây Dựng
Khoa Công Nghệ, Trường Đại Học Cần Thơ

Nội dung chương

1. **Phép đo các vị trí trung tâm** (Measures of Central Location)
2. **Phép đo các biến động** (Measures of Variability)
3. **Qui tắc thực nghiệm**
4. **Vị trí tương đối** (Measures of Relative Standing)
5. **Biểu đồ hộp** (Box Plot)
6. **Phép đo dữ liệu nhóm** (Approximating Descriptive Measures for grouped Data)
7. **Phép đo sự liên hợp** (Measures of Association)

Chương 4:

PHÉP ĐO MÔ TẢ SỐ

NUMERICAL DESCRIPTIVE MEASURES

1. Phép đo các vị trí trung tâm

Measures of Central Location

Đặng Thế Gia, BM Kỹ thuật xây dựng. ĐH Cần Thơ

- **Thông thường chúng ta tập trung mối quan tâm vào hai vấn đề của phép đo các vị trí trung tâm:**
  - Đo điểm trung tâm của dữ liệu (trung bình).
  - Đo sự phân tán (dispersion) của dữ liệu quanh giá trị trung bình.

Điểm trung tâm của dữ liệu phản ánh vị trí của tất cả các điểm dữ liệu thực tế.

- **Trung bình số học (Arithmetic Mean)**
  - Đây là phép đo vị trí trung tâm phổ biến nhất

$$\text{Mean} = \frac{\text{Sum of the measurements}}{\text{Number of measurements}}$$

TB mẫu

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

Kích thước mẫu

TB tổng thể

$$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$$

Kích thước tổng thể

- **Thông thường chúng ta tập trung mối quan tâm vào hai vấn đề của phép đo các vị trí trung tâm:**
  - Đo điểm trung tâm của dữ liệu (trung bình).
  - Đo sự phân tán (dispersion) của dữ liệu quanh giá trị trung bì

Nhưng nếu dữ liệu thứ ba xuất hiện phía trái, nó sẽ "kéo" điểm trung tâm về bên trái.

Nếu dữ liệu thứ ba nằm ngay vị trí trung tâm, điểm trung tâm sẽ không thay đổi

Với 1 điểm dữ liệu, điểm trung tâm nằm ngay vị trí dữ liệu

Với 2 dữ liệu, điểm trung tâm sẽ nằm vị trí giữa (nhằm phản ánh vị trí của cả hai điểm dữ liệu).

- **Ví dụ 1**

Trung bình của mẫu có 6 dữ liệu 7, 3, 9, -2, 4, 6 được tính bởi

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{7 + 3 + 9 - 2 + 4 + 6}{6} = 4.5$$

- **Ví dụ 2**

Giả sử có một hóa đơn tiền điện (tổng thể). Trung bình tổng thể là

$$\mu = \frac{\sum_{i=1}^{200} x_i}{200} = \frac{42.19 + 15.30 + \ldots + 53.21}{200} = 43.59$$

- **Ví dụ 3**

Khi nhiều dữ liệu có cùng giá trị, các dữ liệu có thể được gộp lại thành bảng tần suất.
Giả sử số lao động trẻ em trong một nhóm lao động (mẫu) gồm 16 (kích thước) người như sau:

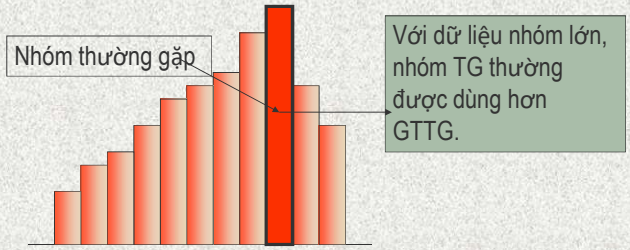
SỐ TRẺ EM	0	1	2	3
SỐ LAO ĐỘNG	3	4	7	2

16 người lao động

$$\bar{x} = \frac{\sum_{i=1}^{16} x_i}{16} = \frac{x_1 + x_2 ... + x_{16}}{16} = \frac{3(0) + 4(1) + 7(2) + 2(3)}{16} = 1.5$$

Giá trị thường gặp (Mode)

- Giá trị thường gặp là giá trị suất hiện với tần suất lớn nhất (xuất hiện nhiều lần nhất).
- Nhóm dữ liệu có thể có một GTTG (hoặc nhóm TG), hoặc nhiều GTTG.



Trung vị (Median)

- Trung vị của một nhóm dữ liệu là giá trị nằm giữa khi dữ liệu được sắp xếp theo thứ tự độ lớn.

Ví dụ 4

Lương của 7 người lao động (đơn vị triệu đồng): **28, 60, 26, 32, 30, 26, 29**. Tìm trung vị của lương

Số lượt quan sát là số lẽ

Trước tiên, xếp lương theo thứ tự tăng dần
Sau đó tìm giá trị nằm chính giữa

26,26,28,29,30,32,60

Giả sử một người lao động nhận lương **31** triệu VNĐ được thêm vào nhóm trên. Tìm trung vị của lương.

Số lượt quan sát là số chẵn

Trước tiên, xếp lương
Sau đó tìm giá trị nằm chính giữa

26,26,28,29, 29.5,30,31,32,60

Có 2 giá trị nằm giữa!

- **Ví dụ 5**
  - Nhà quản lý của của hiệu quấn án nam quan sát thấy size của những thắt lưng (inches) được bán ngày hôm qua là: 31, 34, 36, 33, 28, 34, 30, 34, 32, 40.
  - Giá trị thường gặp của nhóm dữ liệu là 34 in.

Thông tin này có vẽ hữu ích (ví dụ, cho trường hợp thiết kế mới hoặc nhập thêm hàng về kho), hơn là giá trị trung vị 33.5 hay giá trị bình quân 33.2

- **Ví dụ 6**

Thầy giáo dạy môn TKUD muốn báo cáo kết quả thi giữa kỳ của 100 sinh viên. Số liệu như trong bảng sau (file XM04-06). Tìm giá trị bình quân, trung vị, & GTTG? cho biết chúng mô tả thông tin gì?

Marks	
Mean	**73.98**
Standard Error	2.1502163
Median	**81**
Mode	**84**
Standard Deviation	21.502163
Sample Variance	462.34303
Kurtosis	0.3936606
Skewness	-1.073098
Range	89
Minimum	11
Maximum	100
Sum	7398
Count	100

Giá trị bình quân cung cấp thông tin về trình độ tổng thể của lớp. Có thể xem như một công cụ để so sánh với

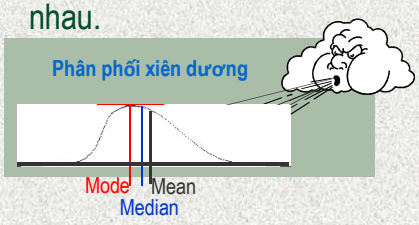
Trung vị chỉ ra rằng có ½ số sinh viên dưới điểm 81 và ½ số sinh viên đạt trên 81.

Kết quả Excel

ho dữ liệu chất
ng chữ (A,B,C,...),
tần suất mỗi điểm có thể được tính toán.
Khi đó GTTG là phép đo hợp lý.

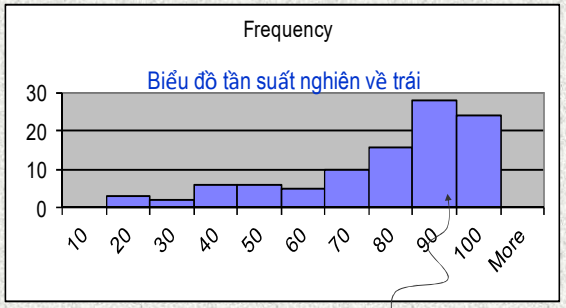
- **Mối quan hệ giữa Mean, Median, và Mode**
  - Nếu một phân phối đối xứng, mean, median và mode sẽ trùng nhau

  - Nếu một phân phối bất đối xứng, và nghiêng (độ xiên) về trái hay phải, 3 giá trị trên sẽ khác nhau.



- **Biểu đồ tần suất Excel (Histogram)**

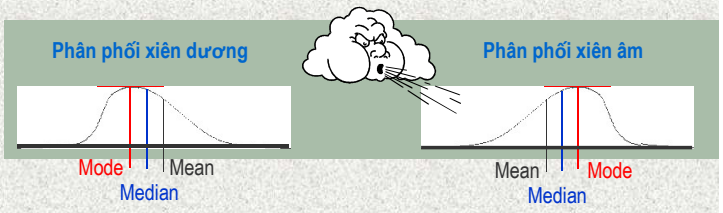
Bin	*Frequency*
10	0
20	3
30	2
40	6
50	6
60	5
70	10
80	16
90	28
100	24
More	0



Nhóm thường gặp (Modal class)

- **Mối quan hệ giữa Mean, Median và Mode**
  - Nếu một phân phối đối xứng, mean, median và mode sẽ trùng nhau

  - Nếu một phân phối bất đối xứng, và nghiêng về trái hay phải, 3 giá trị trên sẽ khác nhau.



Bình quân hình học

- Đây là p… …ng (average growth r…

$$R_g = \sqrt[n]{(1+R_1)(1+R_2)...(1+R_n)} - 1$$

- Gọi R_i là suất thu lợi (R_1, R_2) trong năm i (i=1,2…,n). Bình quân hình học của các năm R_1, R_2, …,R_n là hằng số R_g c… …ứng đến cuối năm thứ n.

R_g được chọn sao cho n giai đoạn sẽ cho cùng kết quả.

Suất thu lợi của n năm được xác định bằng công thức

$$(1+R_1)(1+R_2)...(1+R_n)$$

=

Nếu suất thu lợi là R_g cho tất cả các năm, suất thu lợi trung bình sẽ được tính bởi $(1+R_g)^n$

$$(1+R_g)^n$$

2. Phép đo các biến động

(Nhìn xa hơn giá trị bình quân)

Measures of Variability

(Look beyond the average)

- **Ví dụ 7**
  - Doanh thu 3 năm trước của một doanh nghiệp là $1,000,000
  - Doanh thu tăng hàng năm 20%, 10%, -5%.
  - Tìm bình quân hình học mức tăng của doanh thu.
- **Giải**
  - Gọi R_g là bình quân hình học

$$(1+R)^3 = (1+.2)(1+.1)(1-.05) = 1.2540$$

Vì vậy,

$$R_g = \sqrt[3]{(1+.2)(1+.1)(1-.05)} - 1 = .0784, \text{ or } 7.84\%.$$

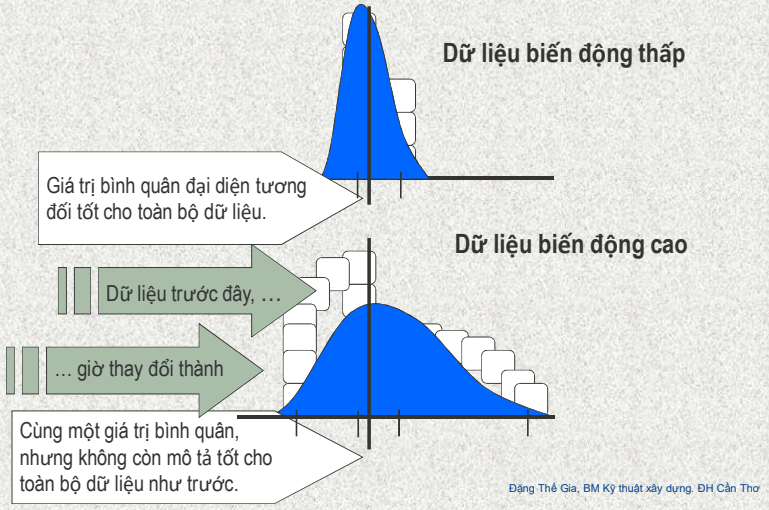
- **Các phép đo vị trí trung tâm không mô tả được toàn bộ câu chuyện về phân phối.**
- **Vẫn còn những thác mắc chưa được trả lời:**

Điển hình của giá trị bình quân của toàn bộ dữ liệu sẽ như thế nào?

hoặc là

Dự liệu trải rộng bao xa quanh giá trị bình quân?

Quan sát hai bộ dữ liệu



▪ Phương sai/Độ lệch quân phương

- Phép đo phân tán này phản ánh giá trị của tất cả các số liệu.
- Phương sai của một **tổng thể** của N số liệu $x_1, x_2,\ldots,x_N$ có giá trị bình quân μ được xác định bằng

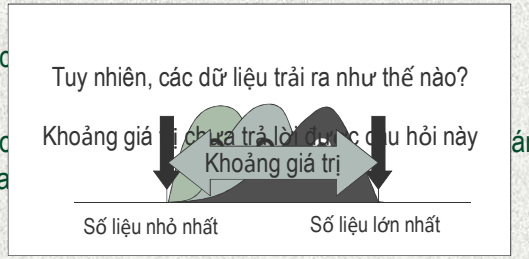
$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

- Phương sai của một **mẫu** của n số liệu $x_1, x_2, \ldots,x_n$ có giá trị bình quân $\bar{x}$ được xác định bằng

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

▪ Khoảng giá trị

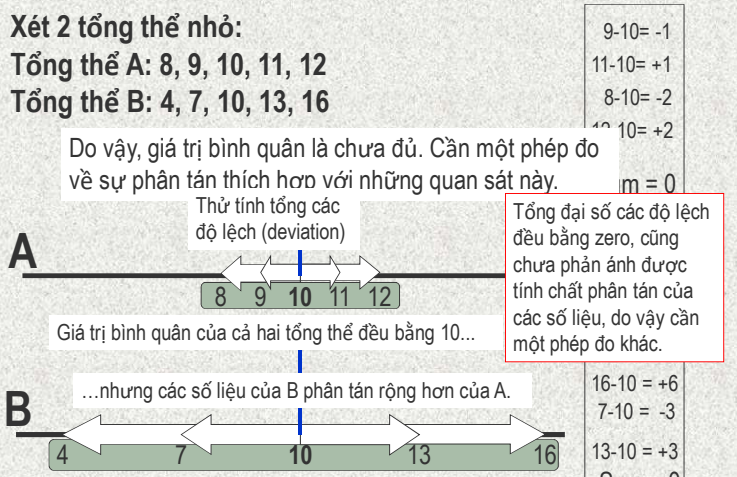
- Khoảng giá trị của bộ dữ liệu là sự chênh lệch của giá trị lớn nhất và giá trị nhỏ nhất.
- Xác
- Kho
của



án

Xét 2 tổng thể nhỏ:
Tổng thể A: 8, 9, 10, 11, 12
Tổng thể B: 4, 7, 10, 13, 16

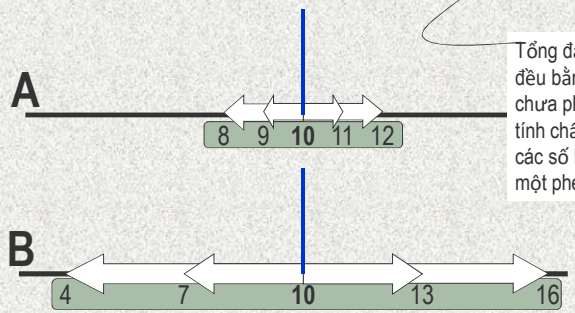
Do vậy, giá trị bình quân là chưa đủ. Cần một phép đo về sự phân tán thích hợp với những quan sát này.



Tổng bình phương các độ lệch là phù hợp để mô tả sự khác biệt.

9-10= -1
11-10= +1
8-10= -2
12-10= +2
Sum = 0

Tổng đại số các độ lệch đều bằng zero, cũng chưa phản ánh được tính chất phân tán của các số liệu, do vậy cần một phép đo khác.



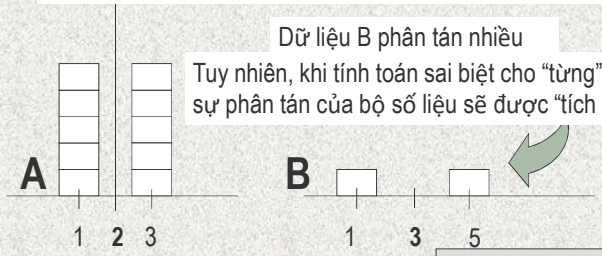
16-10 = +6
7-10 = -3
13-10 = +3
Sum = 0

Bộ dữ liệu nào phân tán nhiều hơn?

Ta tính tổng bình phương các độ lệch cho cả 2 bộ dữ liệu

Dữ liệu B phân tán nhiều

Tuy nhiên, khi tính toán sai biệt cho "từng" số liệu, sự phân tán của bộ số liệu sẽ được "tích lũy".



$Sum_A = (1-2)^2 + \ldots + (1-2)^2 + (3-2)^2 + \ldots + (3-2)^2 = 10$ (5 lần; 5 lần)

$Sum_B = (1-3)^2 + (5-3)^2 = 8$!

$\sigma_A^2 = Sum_A/N = 10/10 = 1$

$\sigma_B^2 = Sum_B/N = 8/2 = 4$

Thử tính phương sai của hai tổng thể A & B

$$\sigma_A^2 = \frac{(8-10)^2+(9-10)^2+(10-10)^2+(11-10)^2+(12-10)^2}{5} = 2$$

$$\sigma_B^2 = \frac{(4-10)^2+(7-10)^2+(10-10)^2+(13-10)^2+(16-10)^2}{5} = 18$$

Tại sao phương sai được định nghĩa là **giá trị bình quân** của bình phương các độ lệch? Tại sao không dùng giá trị **tổng** bình phương?

Còn nữa, tổng bình phương các độ lệch tăng giá trị khi sự phân tán của nhóm dữ liệu tăng lên!!

- **Ví dụ 8**
  - Tìm giá trị bình quân, trung vị, GTTG và phương sai của dữ liệu mẫu sau (đơn vị: năm).

3.4, 2.5, 4.1, 1.2, 2.8, 3.7

- **Giải**

Công thức rút gọn

$$\bar{x} = \frac{\sum_{i=1}^{6} x_i}{6} = \frac{3.4+2.5+4.1+1.2+2.8+3.7}{6} = \frac{17.7}{6} = 2.95$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} = \frac{1}{n-1}\left[\sum_{i=1}^{n} x_i^2 - \frac{(\sum_{i=1}^{n} x_i)^2}{n}\right] =$$

$$= [3.4^2+2.5^2+\ldots+3.7^2] - [(17.7)^2/6] = 1.075 \text{ (năm)}^2$$

- **Độ lệch tiêu chuẩn** (Standard Deviation) của dữ liệu là căn bậc hai của phương sai.

$$Đô\ lêch\ quân\ phuong\ mãu: s = \sqrt{s^2}$$

$$Đô\ lêch\ quân\ phuong\ tông\ thê: \sigma = \sqrt{\sigma^2}$$

Hệ số biến thiên (Coefficient of Variation)

- **Hệ số biến thiên** (CV), còn gọi là **Độ lệch chuẩn tương đối** (Relative SD, RSD) là một đại lượng thống kê mô tả dùng để đo mức độ biến động của tương đối của những tập hợp dữ liệu chưa phân tổ có giá trị bình quân khác nhau.
- Hệ số biến thiên là tỷ số của độ lệch chuẩn và giá trị bình quâ

Độ lệch chuẩn bằng 10 có thể xem là lớn khi giá trị bình quân là 100, nhưng chỉ được xem là vừa phải khi giá trị bình quân là 500

- Hệ số CV tỷ lệ với mức độ biến động của dữ liệu. Dùng để:
  - So sánh độ phân tán giữa các hiện tượng có đơn vị tính khác nhau
  - Hoặc giữa các hiện tượng cùng loại nhưng có số trung bình không bằng nhau.

- **Ví dụ 9**

  Suất thu lợi trong 10 năm qua của hai quỹ tương hỗ được cho như bên dưới. Quỹ nào có mức rủi ro cao hơn?

  Quỹ A: 8.3, -6.2, 20.9, -2.7, 33.6, 42.9, 24.4, 5.2, 3.1, 30.05

  Quỹ B: 12.1, -2.8, 6.4, 12.

- **Giải**

  – Bảng tính bên dưới lấy từ MS Excel (file Xm04-10)

Quỹ A được xem là rủi ro hơn vì có độ lệch chuẩn lớn hơn

Quỹ A		*Quỹ B*	
Mean	16	Mean	12
Standard Error	5.295	Standard Error	3.152
Median	14.6	Median	11.75
Mode	#N/A	Mode	#N/A
Standard Deviation	16.74	Standard Deviation	9.969
Sample Variance	280.3	Sample Variance	99.37
Kurtosis	-1.34	Kurtosis	-0.46
Skewness	0.217	Skewness	0.107
Range	49.1	Range	30.6
Minimum	-6.2	Minimum	-2.8
Maximum	42.9	Maximum	27.8
Sum	160	Sum	120
Count	10	Count	10

Hệ số biến thiên (Coefficient of Variation)

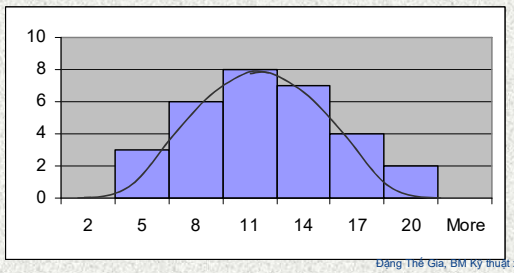
- Giữa 2 tập hợp dữ liệu, tập nào có hệ số biến thiên lớn hơn là tập có mức độ biến động lớn hơn.
- Hệ số biến thiên càng cao, thì độ phân tán của lượng biến càng lớn, tính chất đại diện của số bình quân càng thấp và ngược lại.
- Trong thực tế, thống kê thực nghiệm đã cho rằng nếu CV > 40% tính chất đại biểu của số bình quân thấp.
- Nhược điểm của hệ số biến thiên khi dùng để đo mức độ biến động là nếu giá trị bình quân gần 0 thì chỉ một biến động nhỏ của giá trị bình quân cũng có thể khiến cho hệ số này thay đổi lớn.

3. Qui tắc thực nghiệm

- **Ví dụ 10**
  - Thời gian của 30 cuộc gọi đường dài được mô tả như hình vẽ. Kiểm tra quy tắc thực nghiệm.

  - **Giải**

    Trước tiên kiểm tra liệu biểu đồ tần suất có dạng hình chuông!



- **Độ lệch chuẩn được dùng để**
  - So sánh độ biến động của các phân phối khác nhau
  - Mô tả hình dạng tổng quát của một phân phối
- **Quy tắc thực nghiệm:** Nếu một mẫu số liệu có phân phối dạng hình chuông (gò), khoảng giá trị

$(\bar{x}-s, \bar{x}+s)$ chúa khoang 68% sô liêu
$(\bar{x}-2s, \bar{x}+2s)$ chúa khoang 95% sô liêu
$(\bar{x}-3s, \bar{x}+3s)$ chúa hâu nhu toàn bô sô liêu (99.7%)

- Tính giá trị bình quân và độ lệch chuẩn:
  Mean = 10.26; SD = 4.29.
- Kiểm tra các khoảng:

$(\bar{x}-s, \bar{x}+s) = (10.26 - 4.29,\ 10.26 + 4.29) = (5.97,\ 14.55)$

$(\bar{x}-2s, \bar{x}+2s) = (1.68,\ 18.84)$

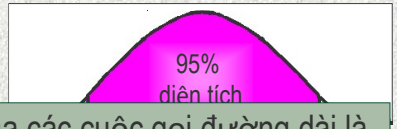
$(\bar{x}-3s, \bar{x}+3s) = (-2.61,\ 23.13)$

Khoảng	Quy tắc TN	Phần trăm xuất hiện
5.97, 14.55	68%	70%
1.68, 18.84	95%	96.7%
-2.61, 23.13	99.7%	100%

Kết luận khác

- Theo quy tắc thực nghiệm, khoảng 95% diện tích phía dưới hình chuông nằm trong khoảng

$(\bar{x} - 2s, \bar{x} + 2s)$

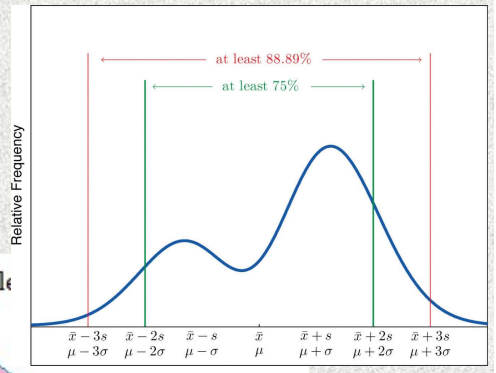


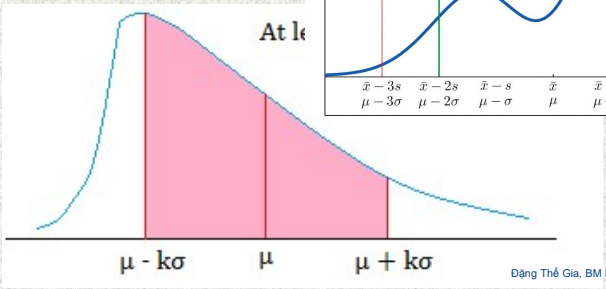
Khoảng giá trị của các cuộc gọi đường dài là 19.5-2.3=17.2 phút

tính gần đúng S

$$s \cong \frac{17.2}{4} = 4.3\,phút \qquad s \cong \frac{KhoangGiáTri}{4}$$

Định lý Chebyshev ($\mathbf{1\text{-}1/k^2}$) đúng cho mọi tập dữ liệu với mọi hình dạng phân phối.





Định lý Chebyshev (theorem)

- Cho một bộ dữ liệu bất kỳ và một số **k** (không nhỏ hơn 1), tỉ lệ dữ liệu nằm trong khoảng **k** lần độ lệch chuẩn quanh Mean tối thiểu là $\mathbf{1\text{-}1/k^2}$.
- Định lý này đúng cho mọi tập dữ liệu với mọi hình dạng phân phối.

$1\text{-}1/2^2=3/4$

$1\text{-}1/3^2=8/9$

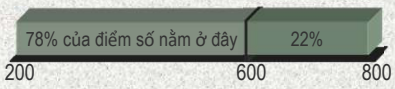
K	Khoảng	Chebyshev	Quy tắc TN
1	$\bar{x} - s, \bar{x} + s$	tối thiểu 0%	xấp xỉ 68%
2	$\bar{x} - 2s, \bar{x} + 2s$	tối thiểu 75%	xấp xỉ 95%
3	$\bar{x} - 3s, \bar{x} + 3s$	tối thiểu 89%	xấp xỉ 99.7%

4. Vị trí tương đối

Measures of Relative Standing

Phân vị

- Phân vị p_{th} của bộ dữ liệu là giá trị tại đó
  - Không quá p% của các dữ liệu nhỏ hơn giá trị đó
  - Không quá (1-p)% của tất cả dữ liệu lớn hơn giá trị đó.
- Ví dụ
  - Giả sử 600 là phân vi 78% của điểm GMAT. Khi đó



- Phân vị 50%, còn gọi là Tứ Phân Vị thứ nhì, chính là **số trung vị (Median)**

- Giải
  - Xếp các số liệu theo thứ tự

Tứ phân vị thứ ba

2, 4, 4, 5, 7, 8, 10, 12, 17, 18, 18, 21, 27, 29, 30

15 số liệu

Tứ phân vị thứ nhất

Tối đa (.25)(15) = 3.75 số liệu nằm dưới Q1. Để ý 3 số liệu đầu tiên ở phía trái.

Không quá (.75)(15)=11.25 số liệu nằm trên Q1. Để ý các số liệu phía phải.

Nếu số số liệu là chẵn, sẽ có hai số liệu để cân nhắc xem số liệu nào là Q1. Khi đó chọn trung bình của hai số liệu này.

- Phân vị thông dụng
  - Thập phân vị thứ nhất (First [lower]decile) = 10%
  - Tứ phân vị thứ nhất (First [lower]quartile, $\mathbf{Q_1}$) = 25%
  - Tứ phân vị thứ nhì (Second [middle]quartile,$\mathbf{Q_2}$) = 50%
  - Tứ phân vị thứ ba (Third [upper]quartile, $\mathbf{Q_3}$) = 75%
  - Thập phân vị thứ chín (Ninth [upper]decile) = 90%

- **Ví dụ 11**

  Tìm tứ phân vị của tập dữ liệu sau

  7, 18, 12, 17, 29, 18, 4, 27, 30, 2, 4, 10, 21, 5, 8

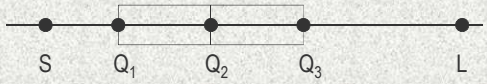
5. Biểu đồ hộp

Box Plot

- Là dạng mô tả bằng hình cho các phép đo mô tả chủ yếu của tập số liệu
  - L - giá trị lớn nhất của số liệu
  - Q3 - tứ phân vị trên
  - Q2 - trung vị
  - Q1 - tứ phân vị dưới
  - S - giá trị nhỏ nhất củ

Khi có các giá trị ngoại biên, cần phải điều chỉnh biểu đồ hộp tổng quát này.

Xem ví dụ phía sau.

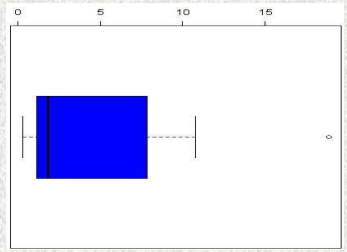


- **Ví dụ 12 – Điều chỉnh khi có giá trị ngoại biên**
  - Ta có bảng số liệu mô tả tỉ lệ CO_2 bình quân đầu người của 8 quốc gia đông dân số nhất thế giới như sau :

Quốc Gia	*CO2/đầu người*
China	4.9
India	1.4
The US	18.9
Indonesia	1.8
Brazil	1.9
Pakistan	0.9
Russia	10.8
Bangladesh	0.3

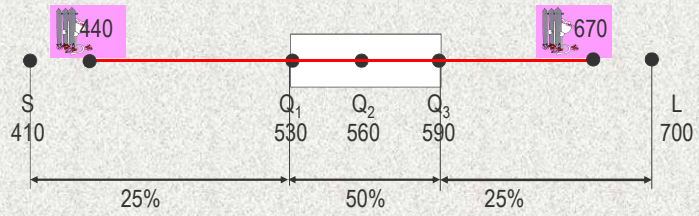
- **Các kiểu "râu" của Biểu đồ hộp**
  - Tối thiểu và tối đa của tất cả các dữ liệu (tổng quát)
  - Mốc thấp nhất vẫn còn trong vòng 1,5*IQR của tứ phân vị dưới, và mốc cao nhất vẫn còn trong vòng 1,5*IQR của tứ phân vị trên (thường được gọi là biểu đồ hộp Tukey, hay John W. Tukey)
  - Một độ lệch chuẩn trên và dưới giá trị bình quân
  - 9% và 91%
  - 2% và 98%

- **Giải**
  - Trước khi vẽ boxplot, ta tính toán các tham số sau:
    - Min = 0.3
    - Q1 = 1.275
    - Trung vị = 1.85
    - Q3 = 6.375
    - Max = 18.9
    - IQR = Q3 – Q1 = 5.1
    - Lower = Q1 – 1.5*IQR = -6.375
    - Upper = Q3 + 1.5*IQR = 14.025

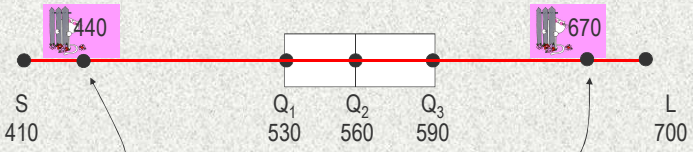


  - Độ trãi giữa (Interquartile Range, IQR = Q3 – Q1)
  - Từ Lower và Upper, ta suy ra US = 18.9 là một giá trị ngoại biên có thể và sẽ không được tính khi vẽ râu của biểu đồ hộp.

- **Ví dụ 13 – điểm GMAT**
  - Vẽ biểu đồ hộp cho dữ liệu về điểm GMAT của 200 sinh viên (file Xm04-12)



- Diễn giải kết quả từ biểu đồ hộp
  - Phổ điểm GMAT trải từ 410 đến 700.
  - Một nửa số điểm thấp hơn 650, và một nửa trên 650.
  - Một nửa số điểm nằm trong khoảng 530 và 590.
  - Một phần tư số điểm thấp hơn 530 và ¼ số điểm trên 590.

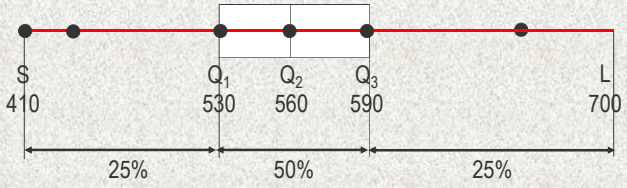


$IQR = Q_3 - Q_1 = 590 - 530 = 60$

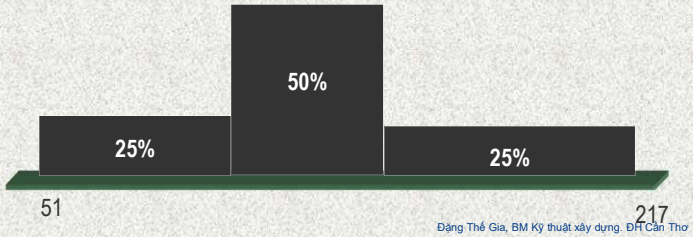
Khoảng trải (Fences) $=\{Q_1-1.5(IQR), Q_3+1.5(IQR\} = \{440, 670\}$

Các giá trị ngoại biên (outliers) là 700 và 410.

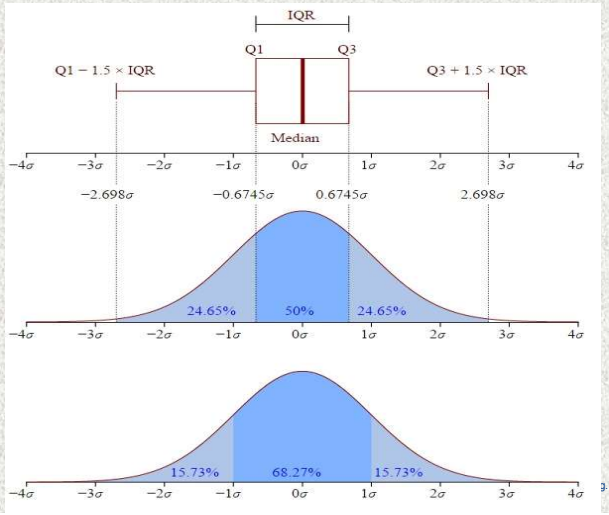
Do vậy, hai "râu" sẽ dời đến 2 ranh giới mới (440, 670), chứ không phải đến giá trị ngoại biên (410 and 700).



Phân phối theo các phân vị là không đối xứng -> Nghiêng dương



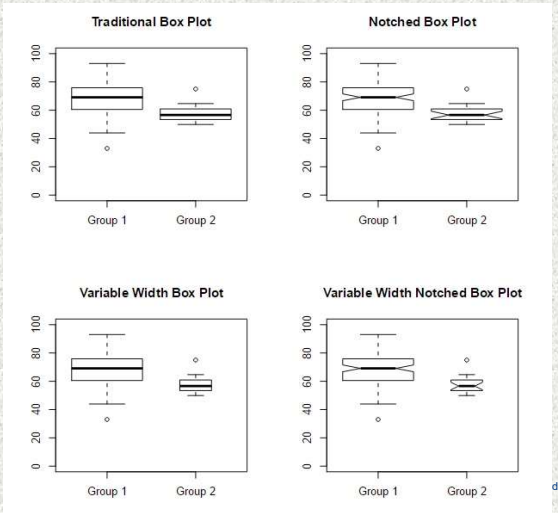
- Các vị trí tương đối của hàm mật độ phân phối chuẩn



6. Phép đo dữ liệu nhóm

Approximating Descriptive Measures for grouped Data

- Biến thể của Biểu đồ hộp



- **Xấp xỉ phép đo mô tả cần thiết trong 2 trường hợp sau:**
  - Khi việc xấp xỉ là cần thiết,
  - Khi chỉ có dữ liệu nhóm thứ cấp.

Số lượng nhóm

$n = f_1+f_2+\ldots+f_k$

Điểm giữa của nhóm i

Tần suất nhóm i

$$\bar{x} = \frac{\sum_{i=1}^{k} f_i m_i}{n}$$

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} f_i m_i^2 - \frac{\left(\sum_{i=1}^{k} f_i m_i\right)^2}{n}\right]$$

$f_i m_i$ là giá trị tương đương xấp xỉ của số liệu nhóm i

- **Ví dụ 14**
  - Xấp xỉ giá trị bình quân và độ lệch chuẩn của độ dài các cuộc gọi từ dữ liệu dạng tần suất

$$\bar{x} \cong \frac{\sum_{i=1}^{6} f_i m_i}{30} = \frac{3120}{6} = 10.4$$

Class i	Class limits	Frequency f_i	Midpoint m_i	$f_i m_i$	$f_i m_i^2$
1	2-5				3.75
2	5-8				53.5
					22.o
		2	18.5	37.0	684.5
		n = 30		312.0	3,751.5

Real values : $\bar{x} = 10.26$ and $s^2 = 18.40$

$$s^2 = \frac{1}{n-1}\left[\sum_{i=1}^{k} f_i m_i^2 - \frac{(\sum_{i=1}^{k} f_i m_i)^2}{n}\right] =$$

$$\frac{1}{29}\left[3{,}751.5 - \frac{312^2}{30}\right] = 17.47$$

- **Hai phép đo mô tả quan hệ tuyến tính giữa hai biến được biểu diễn trên sơ đồ phân tán (scatter diagram).**
  - Hiệp phương sai (Co-variance) – Liệu các biến này biến thiên theo mô hình nào không?
  - Hệ số tương quan (Correlation coefficient) – Quan hệ tuyến tính giữa các biến mạnh như thế nào?

7. Phép đo sự liên hợp

Measures of Association

- **Hiệp phương sai (Co-variance)**

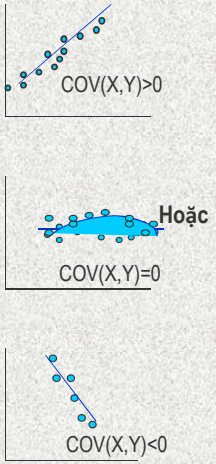
$$\text{Population covariance} = \text{COV}(X,Y) = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$$

μ_x, μ_y là giá trị bình quân của các biến X và Y

N là số phần tử trong tổng thể n là kích thước mẫu.

$$\text{Sample covariance} = \text{cov}(X,Y) = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{n-1}$$

- Nếu hai biến di chuyển theo cùng hướng (cùng tăng hoặc cùng giảm), hiệp phương sai có giá trị dương lớn.
- Nếu hai biến không có quan hệ, hiệp phương sai gần với zero.
- Nếu hai biến di chuyển theo 2 hướng (một tăng, một giảm), hiệp phương sai có giá trị âm lớn.



ρ hoặc r =

+1	Quan hệ tuyến tính dương mạnh
0	Không quan hệ tuyến tính
-1	Quan hệ tuyến tính âm mạnh

- **Hệ số tương quan (coefficient of correlation)**

$$\text{Hệ số tương quan tổng thể}: \quad \rho = \frac{COV(X,Y)}{\sigma_x \sigma_y}$$

$$\text{Hệ số tương quan mẫu}: \quad r = \frac{\text{cov}(X,Y)}{s_x s_y}$$

- Hệ số này trả lời câu hỏi mối quan hệ tuyến tính giữa X và Y mạnh như thế nào.

- Nếu hai biến quan hệ dương mạnh, hệ số tương quan gần với +1 (quan hệ tuyến tính dương mạnh).
- Nếu hai biến quan hệ âm mạnh, hệ số tương quan gần với -1 (quan hệ tuyến tính âm mạnh).
- Không quan hệ theo đường thẳng, hệ số tương quan gần giá trị 0.

Các công thức rút gọn

Công thức

$$\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})=\sum_{i=1}^{n}x_iy_i-\frac{\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n}$$

$$\sum_{i=1}^{n}(x_i-\bar{x})^2=\sum_{i=1}^{n}x_i^2-\frac{\left(\sum_{i=1}^{n}x\right)^2}{n}$$

- Thực hiện các bảng tính bên dưới

Month	x	y	xy	x^2	y^2
1	1	30	30	1	900
2	3	40	120	9	1600
3	5	40	200	25	1600
4	4	50	200	16	2500
5	2	35	70	4	1225
6	5	50	250	25	2500
7	3	35	105	9	1225
8	2	25	50	4	625
Sum	25	305	1025	93	12175

$$\text{cov}(X,Y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n-1}=\frac{1}{n-1}\left[\sum_{i=1}^{n}x_iy_i-\frac{\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n}\right]=\frac{1}{7}\left[1025-\frac{25\times305}{8}\right]=10.268$$

$$s_x^2=\frac{1}{n-1}\left[x_i^2-\frac{\left(\sum_{i=1}^{n}x\right)^2}{n}\right]=\frac{1}{7}\left[93-\frac{23^2}{8}\right]=2.125$$

$$s_x=\sqrt{2.125}=1.458$$

Tương tự, $s_y = 8.839$

$$r=\frac{\text{cov}(X,Y)}{s_xs_y}=\frac{10.268}{1.458\times8.839}=.797$$

- **Ví dụ 15**
  - Tính hiệp phương sai và hệ số tương quan để xem liệu chi phí quảng cáo và doanh thu liên quan với nhau như thế nào?

Advert	Sales
1	30
3	40
5	40
4	50
2	35
5	50
3	35
2	25

Kết quả

	Advertsmn	*sales*
Advertsmnt	2.125	
Sales	10.2679	78.125

Ma trận hiệp phương sai

	Advertsmr	*sales*
Advertsmnt	1	
Sales	0.7969	1

Ma trận hệ số tương quan

Diễn giải

- Hiệp phương sai (10.2679) chỉ ra rằng chi phí quảng cáo và doanh thu quan hệ dương
- Hệ số tương quan (.797) chỉ ra rằng có mối quan hệ tuyến tính dương mạnh giữa quảng cáo và doanh thu.

- **Phương pháp bình phương cực tiểu**
  - Chúng ta tìm một đường thẳng phù hợp nhất với các cặp số liệu
  - Ta định nghĩa "đường phù hợp nhất" là đường có tổng bình phương sai số với các cặp số liệu là tối thiểu.

$$\text{Minimize} \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Giá trị y thực tế của điểm i

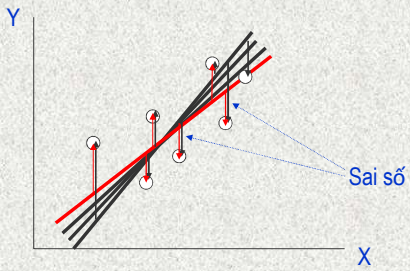
Giá trị y của điểm i được tính từ phương trình

$$\hat{y}_i = b_0 + b_1 x_i$$

Hệ số b_0 và b_1 của đường thẳng làm tối thiểu tổng bình phương của các sai số được tính từ các số liệu

$$b_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad b_0 = \bar{y} - b_1\bar{x}$$

$$với\ \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}\ và\ \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$



Những đường khác nhau cho sai số khác nhau, vì vậy sẽ cho tổng bình phương các sai số khác nhau.

XIN CẢM ƠN!