Chương 2. CÁC ĐẠI LƯỢNG ĐẶC TRƯNG CỦA DÂN SỐ VÀ MẪU

II.1 Dân số và mẫu

- Dân số (population) làm một tập hợp các cá thể có chung một số đặc tính đã được xác định.
- Mẫu (sample): một tiểu tập hợp trong dân số. Phân tích số liệu trên mẫu có thể suy ra các đặc tính cho toàn bộ dân số với độ tin cậy xác định nào đó.

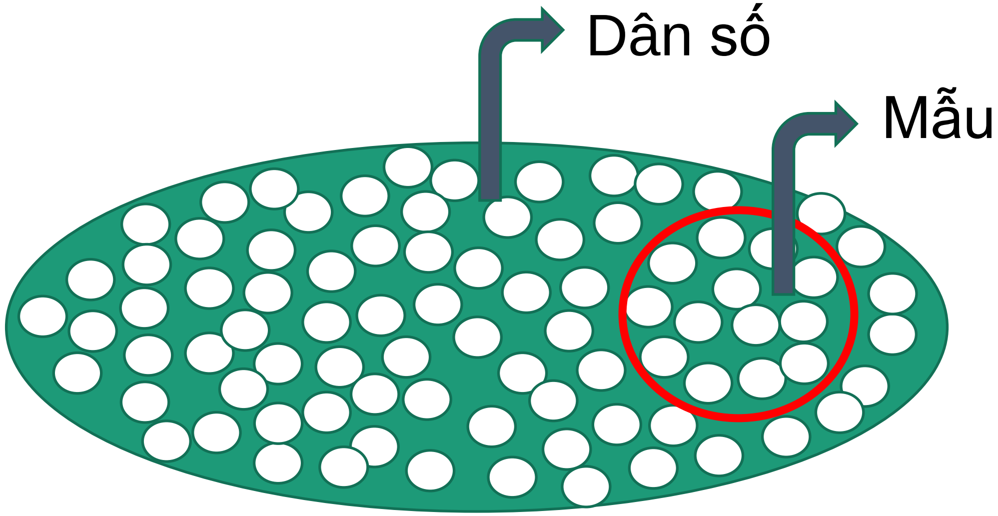
Hai đại lượng đo quan trọng nhất trong thống kê là đo sự tập trung và đo độ phân tán.

Bảng 2.1. Các ví dụ về dân số và mẫu.

Vấn đề	Dân số	Mẫu
Điều tra quan điểm	Toàn bộ sinh viên trong trường	Một số sinh viên trong từng lớp
Chấp nhận hay từ chối lô hàng	Toàn bộ lô hàng	Kiểm tra một số mẫu trong lô hàng
Năng suất trồng rau	Toàn bộ các nông dân trong toàn tỉnh	Một số hộ nông dân trong hợp tác xã

Chương 2. CÁC ĐẠI LƯỢNG ĐẶC TRƯNG CỦA DÂN SỐ VÀ MẪU

2.1 Dân số và mẫu



Đo sự tập trung	Đo độ phân tán
• Số trung bình • Số trung vị • Số thường hiện	• Đo khoảng cách • Đồ thị tần số • Phương sai và độ lệch chuẩn • Sai số chuẩn • Hệ số biến động

Bảng 2.1. Các ví dụ về dân số và mẫu

Vấn đề	Dân số	Mẫu
Điều tra quan điểm	Toàn bộ sinh viên trong trường	Một số sinh viên trong từng lớp
Chấp nhận hay từ chối lô hàng	Toàn bộ lô hàng	Kiểm tra một số mẫu trong lô hàng
Năng suất trồng rau	Toàn bộ các nông dân trong toàn tỉnh	Một số hộ nông dân trong hợp tác xã

2.2 Đo sự tập trung

Các số trong một mẫu có khuynh hướng tập trung về một số nào đó. Các đại lượng đặc trưng cho sự tập trung gồm số trung bình, trung vị và số thường hiện.

2.2.1 *Số trung bình (mean)*

Trung bình của mẫu có n dữ liệu được tính theo nhiều công thức:

- Trung bình số học của từng nghiệm thức (treatment A, B, C):

$$Y_{tbA} = \sum y_n/n_A \qquad Y_{tbB} = \sum y_n/n_B \qquad Y_{tbC} = \sum y_n/n_C \qquad (2.1)$$

- Trung bình số học của cả 3 nghiệm thức:

$$Y_{tbW} = (Y_{tbA} + Y_{tbB} + Y_{tbC}) / 3 \qquad (2.2)$$

- Trung bình trọng lượng của 3 nghiệm thức:

$$Y_{tbWM} = (n_A Y_{tbA} + n_B Y_{tbB} + n_C Y_{tbC}) / (n_A + n_B + n_C) \qquad (2.3)$$

- Trung bình tích:

$$Y_{tb} = (y1.y2\ldots\ldots yn)^{1/n} \qquad (2.4)$$

Ví dụ: Tính chiều dài trung bình của xoài

	Tháng 1	**Tháng 2**	**Tháng 3**
Y_1	12	13	10
Y_2	15	16	14
Y_3	14	17	12
Y_4		15	
Y_{tb} (equ 2.1)			
Y_{tbW} (equ 2.2)			
Y_{tbWM} (equ 2.3)			

2.2.1 *Số trung bình (mean)*

- Trung bình theo tần số của các cá thể (một dạng của trung bình trọng lượng):

$$Y_{tbW} = \sum f_i y_i / \sum f_i \qquad (2.5)$$

Với f_i là tần số của các các thể có cùng đại lượng đo.

Nếu giá trị y_i thu được nằm trong khoảng y_{min} và y_{max} thì giá trị trung bình được lấy làm đại diện cho khoảng đó: $y = (y_{min} + y_{max})/2$.

Ví dụ: đo đường kính của 100 bắp cải

Đường kính (cm)	**Tần số f_i**
20-22	8
22-24	60
24-26	32
y_{tbW}	

Ví dụ: tính năng suất lúa bình quân vụ mùa của một số xã từ số liệu sau:

Y_i (ha/tạ)	f_i (diện tích, ha)	Y_i (đại diện)	f_i Y_i (đại diện)
<30	150		
30-35	100		
35-40	200		
40-45	400		
45-50	250		
>50	50		
Tổng			
y_{tbW}			

2.2.2 *Số trung vị (median)*

Là số nằm giữa dãy số khi dãy số được sắp xếp từ nhỏ đến lớn. Số trung vị cho kết quả nhanh về ước lượng trung bình.

Ví dụ: số bao bì bị ghép mí lỗi trong mỗi lô hàng là 20, 35, 40, 55, 55, 60, 71, 72, 83, 90.

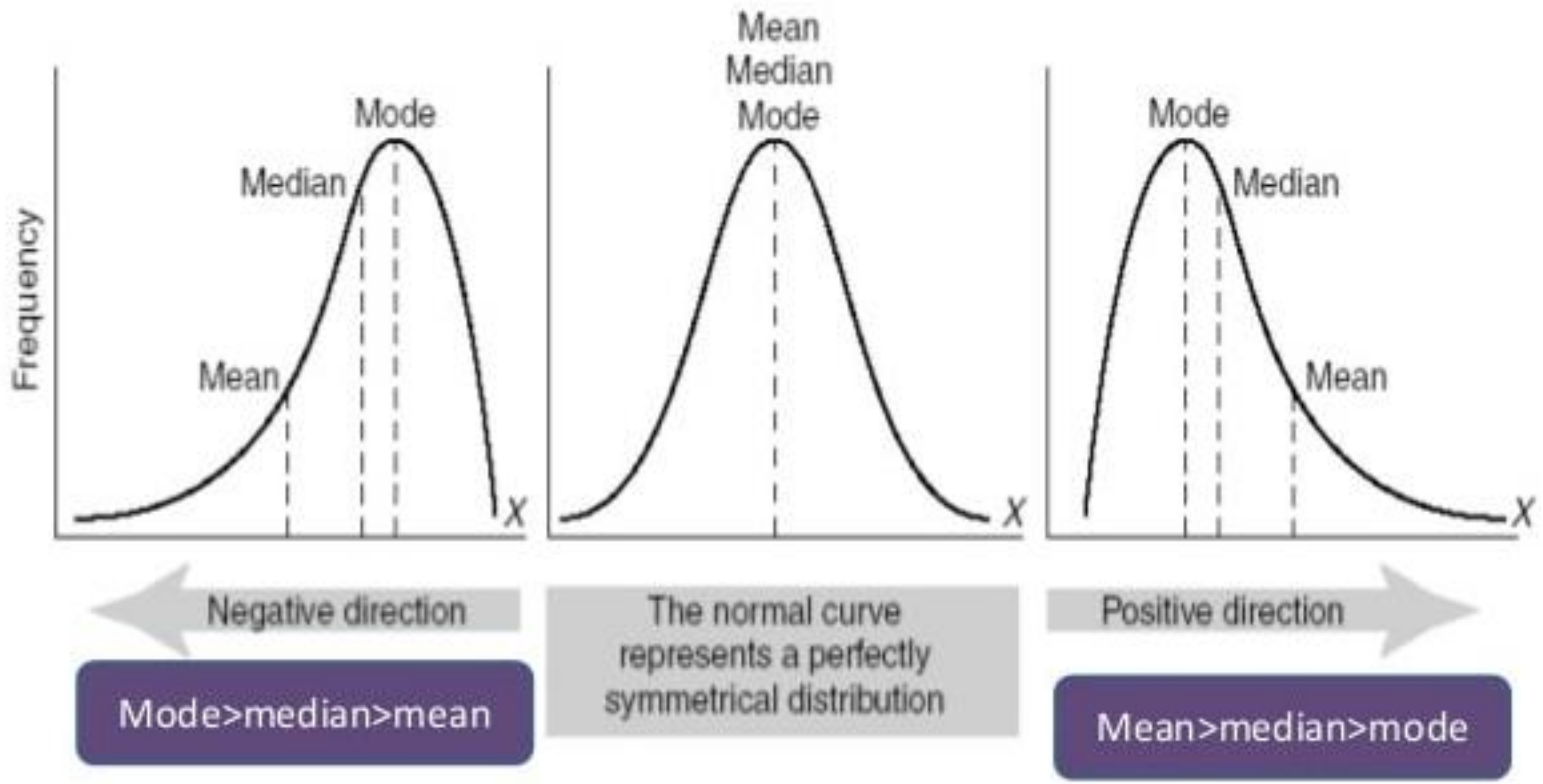
Số trung vị sẽ là ……………………………trong khi số trung bình là ……………………..

2.2.3 *Số thường xuất hiện (mode)*

Là số có tần số lớn nhất.

Có thể có 1 hoặc nhiều

số thường hiện.



2.3 Đo độ phân tán

Là đo mức độ chênh lệch giữa các cá thể của dân số.

2.3.1 Đo khoảng cách

Là số đo khoảng cách giữa số lớn nhất và nhỏ nhất của mẫu. Đại lượng này cho thông tin về khoảng cách của số liệu.

Ví dụ: số bao bì bị ghép mí lỗi trong mỗi lô hàng là 20, 35, 40, 55, 55, 60, 71, 72, 83.

Số đo khoảng cách =

2.3.2 *Đồ thị tần số*

Số liệu được vẽ lên để cho thấy hình ảnh phân bố và sự phân tán của mẫu.

2.3.3. Phương sai và độ lệch chuẩn

- Phương sai (variance): là số đo độ lệch bình phương trung bình của dữ liệu so với số trung bình.

- Phương sai của dân số: N số liệu, trung bình dân số là μ:

$$\sigma^2 = \Sigma (y_i - \mu)^2 / N \quad (2.6a)$$

- Phương sai của mẫu: n số liệu, trung bình mẫu là Y_{tb}:

- $$S^2 = \Sigma (y_i - Y_{tb})^2 / (n-1) = [n \Sigma y_i^2 – (\Sigma y_i)^2] / n(n-1) \quad (2.6b)$$

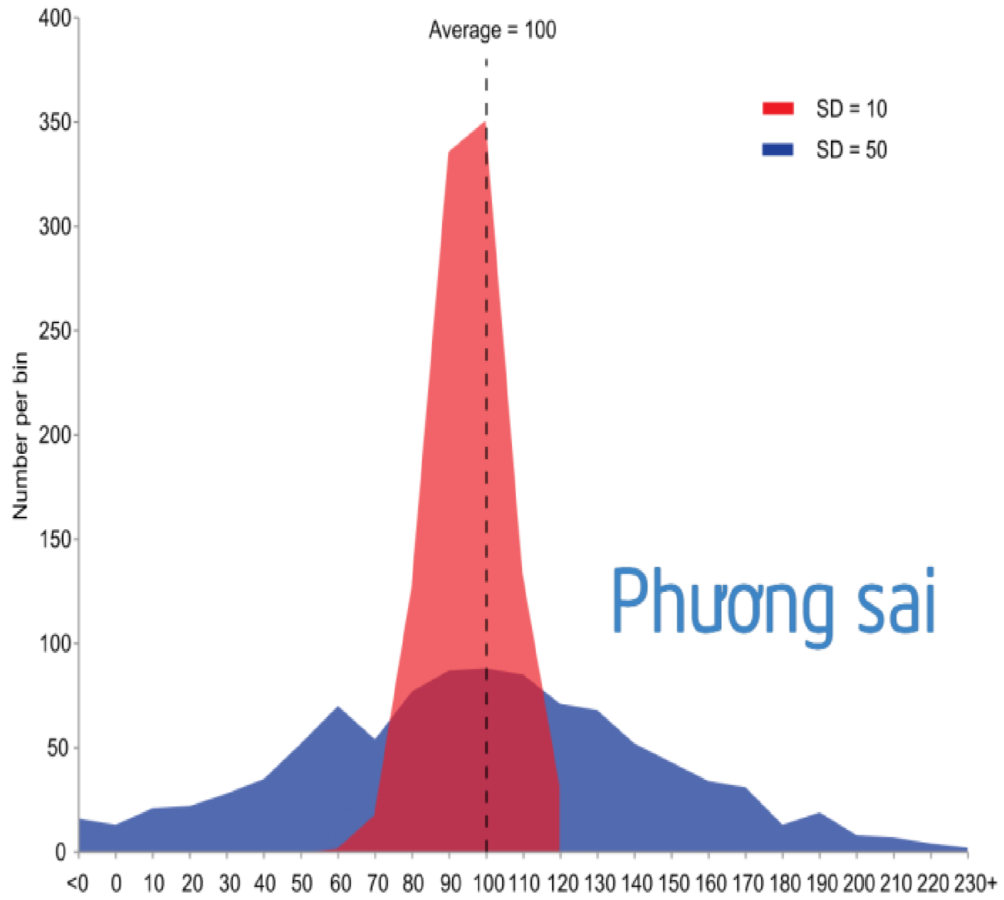
Đối với dân số, độ tự do là N và đối với mẫu độ tự do là n-1 vì phải cần một bậc tự do để đánh giá số trung bình.

- Độ lệch chuẩn (standard deviation): là một đại lượng thống kê dùng để đo mức độ phân tán của một tập hợp dữ liệu đã được lập thành bảng tần số.

- \- Độ lệch chuẩn của dân số: $\sigma = \sqrt{\sigma^2}$

- \- Độ lệch chuẩn của mẫu: $S = \sqrt{S^2}$

- Ý nghĩa: cho thấy sự chênh lệch về giá trị của từng thời điểm đánh giá so với giá trị trung bình



2.3.4. Sai số chuẩn (standard error)

Thực hiện các thí nghiệm khác nhau trên một dân số cho nhiều kết quả trung bình Y_{tb} khác nhau. Sai số chuẩn (SE) là đại lượng dùng để đánh giá khoảng trung bình của dân số (μ) khi biết trung bình của mẫu thí nghiệm.

$$\mu = Y_{tb} \pm t\,(SE) \qquad (2.8a)$$

t là giá trị của tiêu chuẩn Student cho trong bảng phân bố Student (xem mục 2.4 và phụ lục 2.2). Sai số chuẩn được tính như sau:

$$SE = S / \sqrt{n} \qquad (2.8b)$$

2.3.4. *Sai số chuẩn (standard error)*

Ví dụ: Xác định chiều dài thực của trái xoài dựa trên các giá trị sau

Số lượng mẫu n = 30

Độ tin cậy 95%, giá trị t trong bảng phân bố là $t_{(0.05;30-1)} = 2.045$.

Giá trị trung bình mẫu $Y_{tb} = 12$ cm

Phương sai S = 2.5

$\mu = Y_{tb} \pm t\,(SE) =$ ………………………………………………………….

Kết luận: từ thí nghiệm cho số trung bình $Y_{tb} = 12$ cm, chiều dài thực của trái xoài nằm trong khoảng ………….cm đến 12.93 cm với độ tin cậy 95%.

Chú ý: khi số lượng mẫu n >30 thì ta thay t bằng Z là giá trị xác suất của phân bố chuẩn cho sẵn trong bảng (xem mục 2.4).

2.3.5. *<u>Hệ số biến động (coefficient of variation)</u>*

Hệ số biến động (CV) dùng để đánh giá sự biến thiên của S so với số trung bình, nó giúp ta so sánh sự biến thiên giữa hai mẫu độc lập với đơn vị đo lường.

$$CV\ (\%) = 100\ (S/Y_{tb})$$

Ví dụ: Đánh giá chất lượng làm việc của 2 máy đóng gói như sau

Máy A	Máy B
Trọng lượng gói Y_{tbA} = 21 g	Trọng lượng gói Y_{tbB} = 15 g
S_A = 3.2 g	S_B = 3 g
CV_A = ……………………	CV_B = ……………………..

Kết luận: ……………………………………………………………………

2.4 Ước lượng cỡ mẫu

Số lượng của mẫu cần thiết có thể tính toán từ công thức sau:

$$n \geq (S.Z / e)^2$$

$Z = (Y_{tb} - \mu) / SD$: giá trị của phân bố chuẩn có thể lấy từ bảng phân bố chuẩn (phụ lục 2.1)

$e = Y_{tb} - \mu$: khác biệt giữa trung bình mẫu (đại lượng đo được) và trung bình dân số (giá trị thực).

Ví dụ: khi kiểm tra trọng lượng tịnh của sản phẩm sau khi đóng gói cho kết quả độ lệch chuẩn S = 15.6 g. Hỏi số lượng gói cần đo là bao nhiêu để trọng lượng trung bình của mẫu khác biệt so với trọng lượng yêu cầu là nhỏ hơn 5 g?

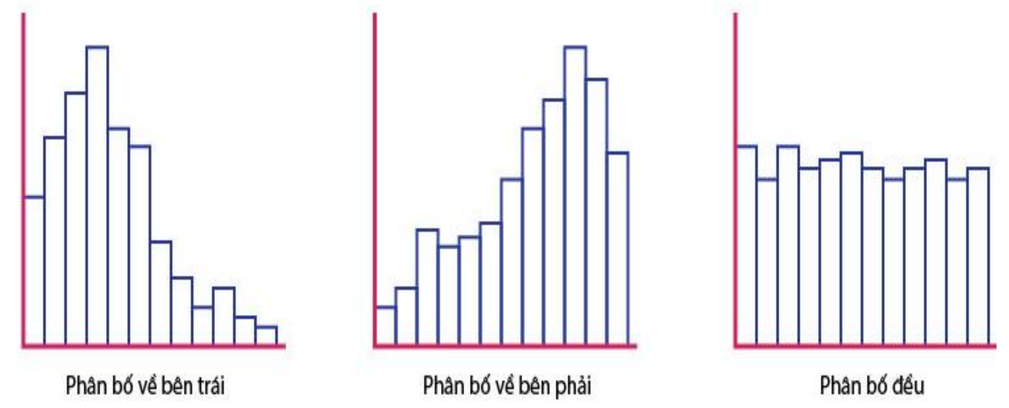
Giải: với độ tin cậy 95%, xác suất P = 0.05/2 = 0.025 nên giá trị Z = 1.96 (từ bảng):

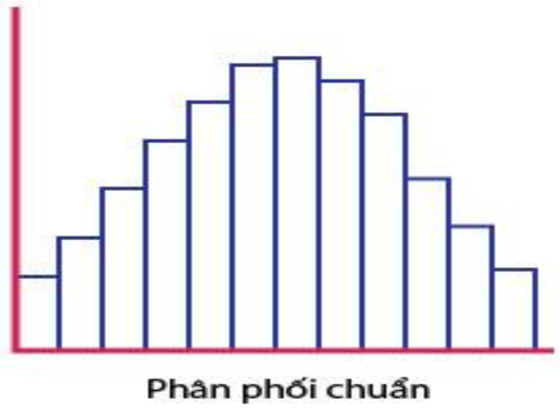
n ≥……………………………………..

Kết luận: số mẫu ít nhất cần lấy để đo trọng lượng tịnh sản phẩm là ……. gói

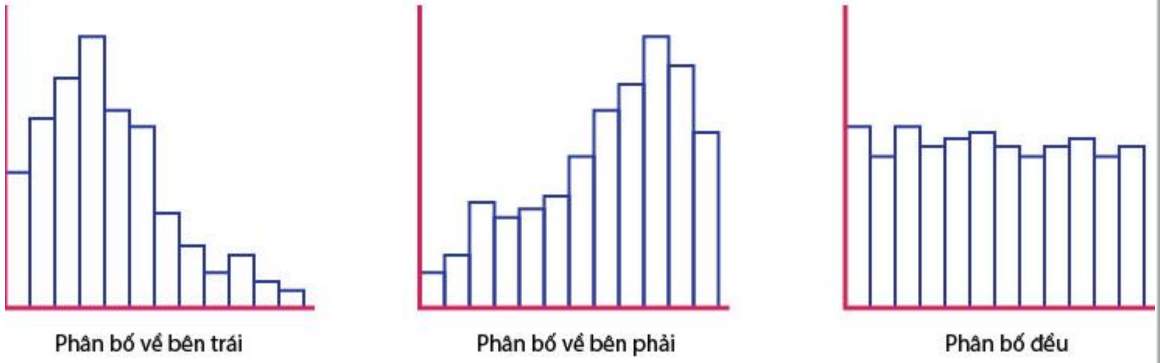
2.5 Các loại phân bố và cách dùng các bảng xác suất

- Khi vẽ biểu đồ tương quan giữa tần số và kết quả thu nhận, hình dạng đường cong thường có dạng hình chuông và gọi là phân bố chuẩn. Tuy nhiên có nhiều trường hợp là phân bố không chuẩn. Lúc này ta có thể dùng phương pháp biến đổi số liệu để chuyển phân bố của số liệu sang phân bố chuẩn



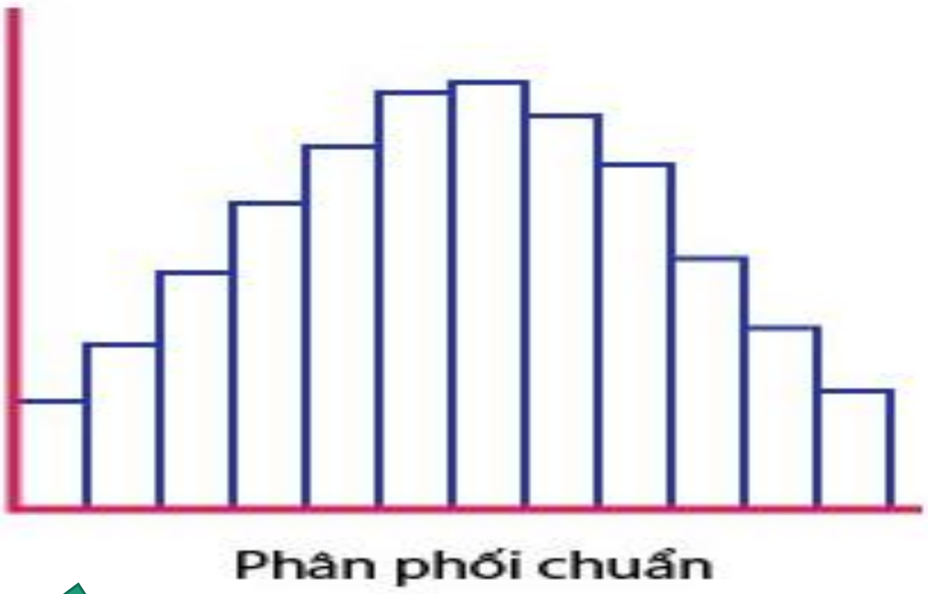


2.5 Các loại phân bố và cách dùng các bảng xác suất



Hình. Các dạng đường cong của biểu đồ tần số có thể gặp

Biến đổi số liệu



2.5.1 Phân bố chuẩn (normal distribution)

Một tập dữ liệu được xem là có phân bố chuẩn thì sẽ có các thuộc tính sau:

- Số trung bình (mean) = số trung vị (median) = số thường hiện (mode)
- Có tính đối xứng (hình chuông)
- 50% giá trị sẽ nhỏ hơn giá trị trung bình và 50% giá trị sẽ lớn hơn giá trị trung bình.
- Công thức của phân bố chuẩn dựa trên 3 đại lượng số trung bình, phương sai và độ lệch chuẩn là:

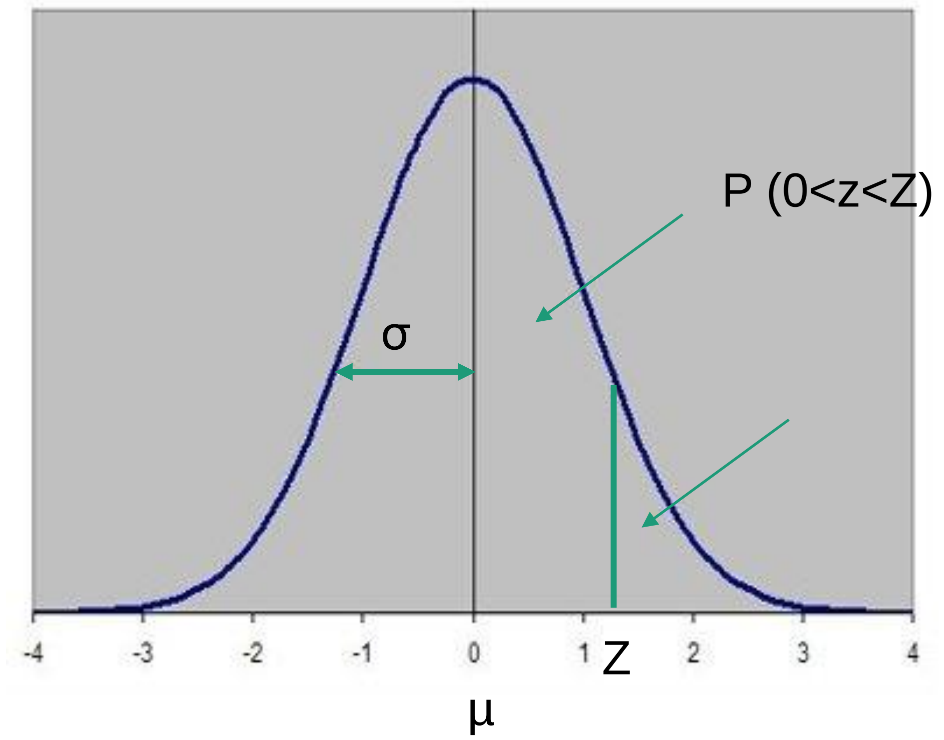
$$P_{(Y=y)} = f(y) = \frac{1}{\sigma\sqrt{2\pi}} e\left[\frac{-(y-\mu)^2}{2\sigma^2}\right] \qquad (2.11)$$

Tần số f(y) của giá trị thu được y có dạng hình chuông biểu diễn bằng phương trình (2.11) gọi là phân bố chuẩn, ký hiệu NID (normal independent distribution). Thông thường biến Z được dùng thay thế biến y để loại bỏ thứ nguyên:

$$Z = (y-\mu) / \sigma \qquad (2.12)$$

2.5.1 <u>Phân bố chuẩn (normal distribution)</u>

- Giá trị trong bảng thường cho với μ = 0 và σ = 1.
- Xác xuất để z nằm trong khoảng 0 đến Z: ký hiệu P (0<z<Z) là diện tích bên dưới đường phân bố từ 0 đến Z
- Xác xuất để z nằm trong khoảng Z đến ∞: ký hiệu P (Z < z < ∞) là diện tích bên dưới đường phân bố từ Z đến ∞.
- Trong thực tế nhiều biến số tuân theo phân bố chuẩn: kích thước, trong lượng, năng suất



Giá trị P (P-value): được đo bởi diện tích giới hạn bởi đường cong chuẩn

Thay đổi μ: dịch chuyển phân bố qua trái hoặc phải

Thay đổi σ: làm tăng hoặc giảm độ phân tán

2.5.1 Phân bố chuẩn (normal distribution)

Ví dụ: Khối lượng tịnh bình quân mì gói sản xuất từ nhà máy là 100 g với phương sai 25.

1. P (100 < y < 110)?

2. y > 110 g có hiếm xảy ra?

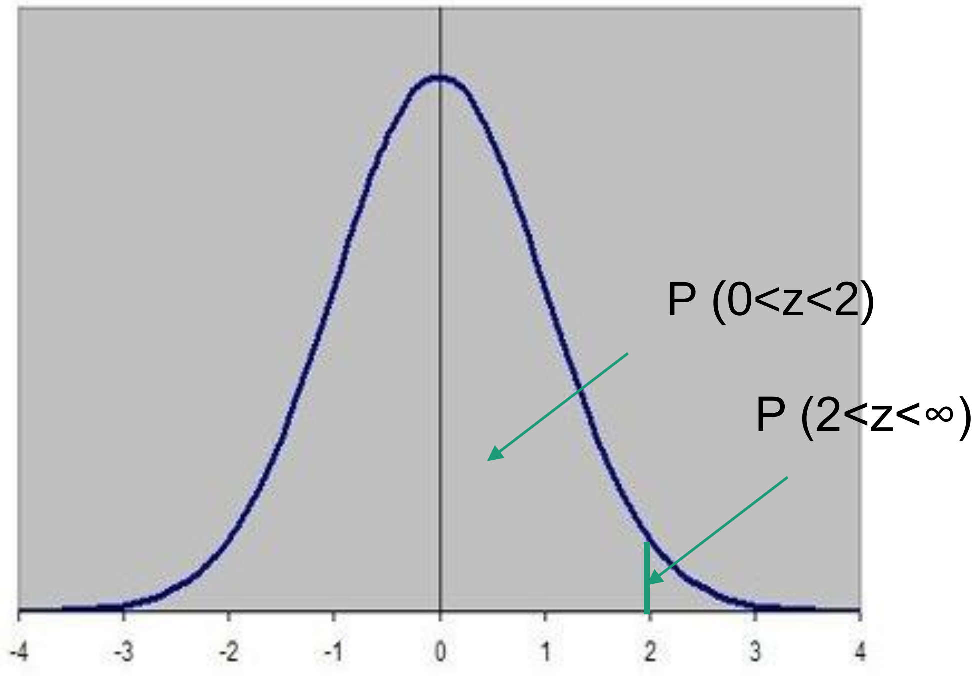
Giải:

Bước 1. Chuyển đổi số liệu sang phân bố chuẩn

…………………………………………………………..

Bước 2. Tra bảng phân bố chuẩn

1. P (100 < y < 110)= ……………………………………



2. P (y >110) = P (110 < y < +∞) =…………………………………………………………..

hoặc P (y < 110) = P (- ∞ < y < 110) =…………………………………………..

2.5.1 Phân bố nhị thức (binomial distribution)

Phân bố nhị thức là một dạng của phân bố rời rạc thường được dùng trong thống kê, ngược lại của các dạnh phân bố liên tục như phân bố chuẩn.

Mỗi quan sát độc lập trong phân bố nhị thức chỉ được phân loại vào một trong hai lớp trong một số lượng lần thử

Ví dụ: nấm mốc sống hay chết sau khi xử lý nhiệt; bao bì ghép mí kín hay hở, trứng thụ tinh hay không, giống đực hay cái.

Phân phối nhị thức xác định xác suất quan sát được một số lần thành công nhất định trong số lần thử nhất định.

.

Phân phối nhị thức thể hiện xác suất để **y** thành công trong **n** phép thử, với xác suất thành công của mỗi phép thử là **p**

Gọi : p= xác suất của kết quả thu nhận y trong n phép thử, q: xác suất thu nhận còn lại

Ta có: p + q = 1. Nếu n: số mẫu;

Giá trị trung bình của phân phối nhị thức: $Y_{tb} = np$

Phương sai của phân phối nhị thức: $S = (npq)^{1/2} = (np(1-p))^{1/2}$

Xác suất **p** để thu nhận được **y** từ **n** mẫu là: $p(y) = C_n^y p^y q^{n-y}$ (2.13)

Trong đó: $C_n^y = \frac{n!}{y!(n-y)!}$ (2.14)

Khi số mẫu n rất lớn (n→∞) và p→0, công thức (2.13) trở thành:

$$p(y) = \lim_{n\to\infty} C_n^y p^y q^{n-y} = \frac{a^x}{x!} e^{-a} \quad (2.15)$$

Trong đó, $Y_{tb} = np = a$; $S_y = (npq)^{1/2} = a^{1/2} = (Y_{tb})^{1/2}$. Đây là phân bố Poisson.

Khi n > 30: phân bố nhị thức gần với phân bố chuẩn. Hoặc khi p=q=0.5, phân bố nhị thức trở thành phân bố chuẩn.

<u>Ví dụ 1</u>: Kiểm tra 500 con cá thì có 427 cá đực. Tính xác suất để thu nhận được số các đực như trên.

p =…………………………………………   q = 1 – p = …………………..

P(427) = ………………………………………………………………………..

Xác suất để thu nhận số cá đực như trên là 5.05%

<u>Ví dụ 2</u>: Tỷ lệ đồ hộp bị hư hỏng sau tiệt trùng là 5%. Tính xác suất trong lô hàng 5000 hộp có 300 hộp bị hư hỏng.

Như vậy y = 300, n = 5000, p = 0.05 và q = 0.95

P(y=300;5000;0.05) = ………………………………………………………………………………………………..

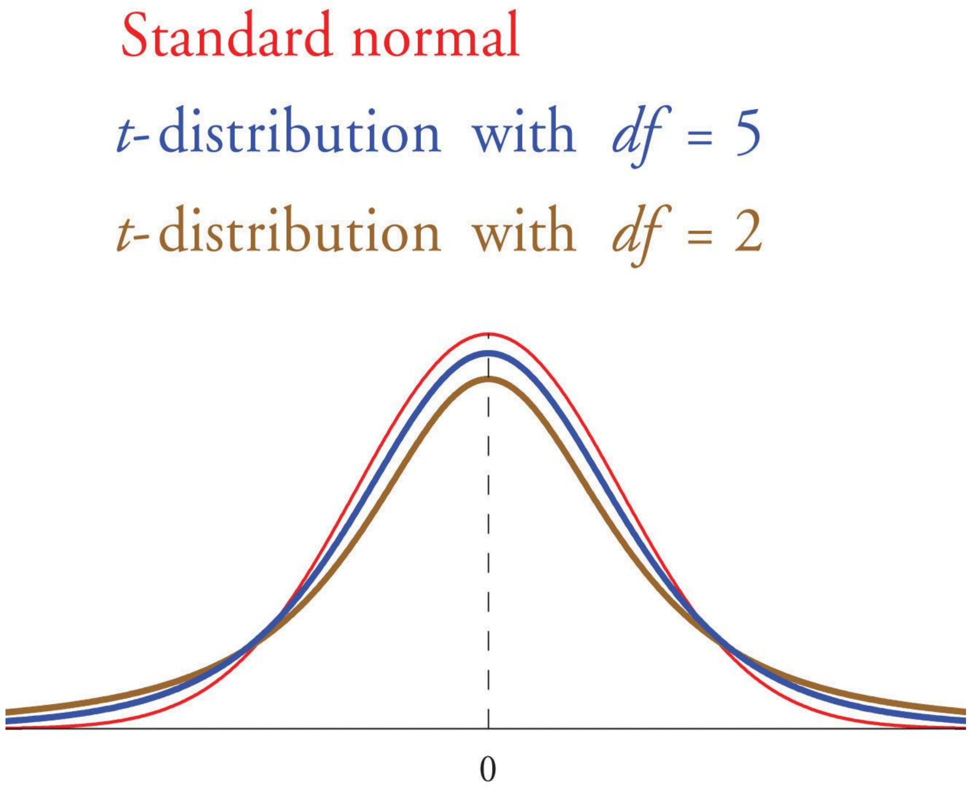
Xác xuất để bị 300 hộp hư hỏng là …………………………………………………

2.5.3 Phân bố Student (phân bố t)

- Phân bố t được dùng trong các trường hợp kích thước mẫu quá nhỏ
- Độ lệch chuẩn σ của dân số không biết.

$$t = \frac{Y_{tb} - \mu}{\frac{S}{\sqrt{n}}}$$

- Hình dạng đối xứng gần giống với phân phối chuẩn
- Khi n > 30, thích hợp nhất là 100, phân bố t gần với phân bố chuẩn.
- Cỡ mẫu càng nhỏ, phần đuôi càng nặng và xa hơn



2.5.3 Phân bố Student (phân bố t)

__Ví dụ:__

1. Với số mẫu n = 30, độ tin cậy 95% (mức ý nghĩa α = 0.05), giá trị t=?

Tra bảng phân bố student: $t_{bảng} = t_{α(n-1)} = t_{0.05(30-1)}$ =…………………

n-1: df: bậc tự do

2. Với số mẫu n = ∞, độ tin cậy 95% (mức ý nghĩa α = 0.05),

$t_{0.05(∞)}$ = ……………………………ở mức ý nghĩa 0.05

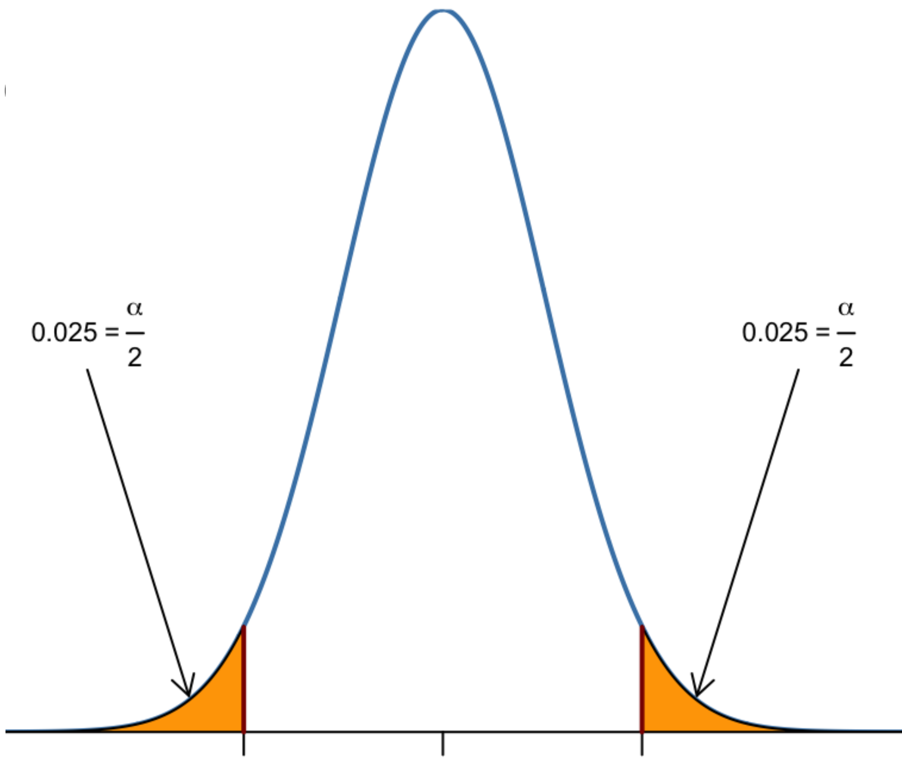
Ở bảng phân bố chuẩn:

Z=1.96 thì P(0 < z < 1.96) = …………………

hay P(-1.96 < z < 19.6) = ……………………..

→mức ý nghĩa α = 0.05 trùng với phân bố t.

Vậy khi số mẫu lớn thì phân bố t trùng với phân bố Z



2.5.3 *Phân bố F*

Cho 2 dân số có phân bố chuẩn có phương sai lần lượt là σ_1^2 và σ_2^2. Ta lấy ngẫu nhiên 2 mẫu có kích thước tương ứng là n_1 và n_2. Các mẫu này có phương sai là S_1^2 và S_2^2, biến số:

$$F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \qquad (2.16)$$

có xác suất phân bố theo qui luật gọi là phân bố F (viết tắt của nhà thống kê R.H.Fisher)

Muốn đọc giá trị F (phụ lục 4.3) thì cần biết mức ý nghĩa α, độ tự do $v_1 = (n_1 – 1)$ của dân số thứ 1 và $v_2 = (n_2 – 1)$ của mẫu số (dân số thứ 2).

Các bảng tra cứu phân vị F chỉ tương ứng với các giá trị thấp của α như 0.1 hay 0.05. Trong trường hợp α có giá trị lớn hơn như 0.95 thì dùng công thức:

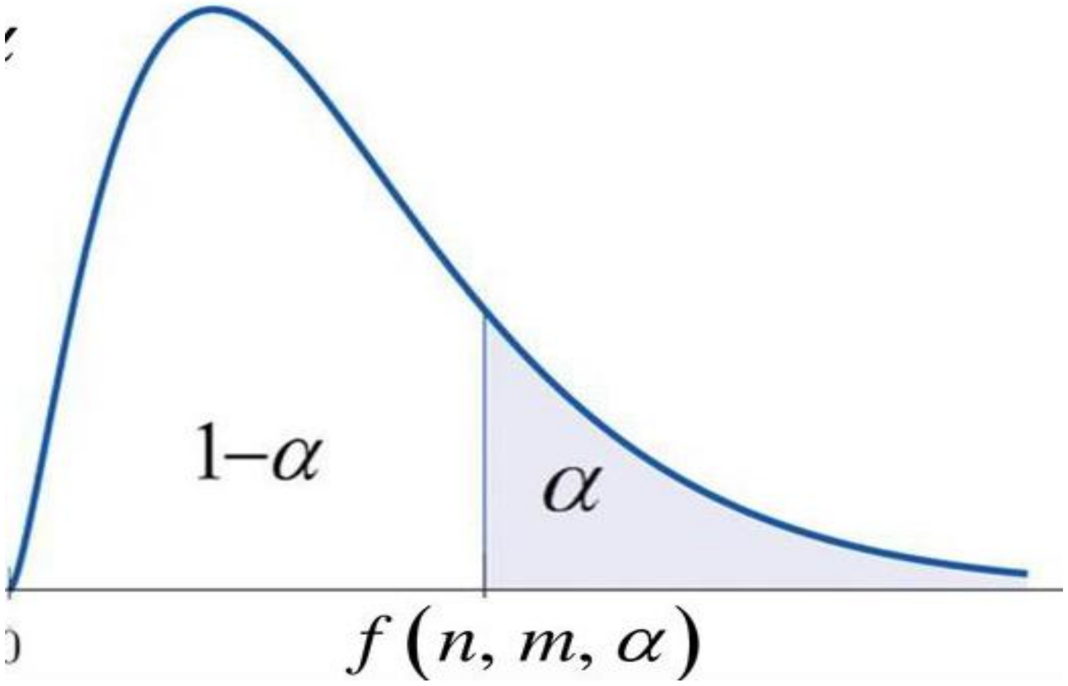
$$F_{1-\alpha,\, v1,\, v2} = \frac{1}{F_{\alpha, v1, v2}}$$

Ví dụ: $F_{0.05(4,20)}$ = $F_{0.05(20,4)}$ = $F_{0.95(20,4)}$ =

Giá trị tới hạn mức α của phân phối F (v1;v2), ký hiệu là $f_\alpha(v1,v2)$, được xác định qua đẳng thức:

$$P\,(F > f_{\alpha}(n1\,,n2)) = \alpha$$

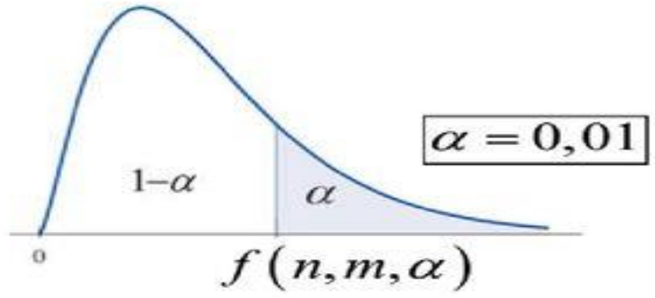
Tính chất: $F_{1-\alpha,\, v1,\, v2} = \dfrac{1}{F_{\alpha,v1,v2}}$



Để tra giá trị F (phụ lục 4.3): cần mức ý nghĩa α, độ tự do $v_1 = (n_1 - 1)$ và $v_2 = (n_2 - 1)$

Ví dụ: $F_{0.05(4,15)}$ = $F_{0.05(15,4)}$ = $F_{0.95(15,4)}$ =

Bảng giá trị tới hạn Fisher



PHỤ LỤC 4. GIÁ TRỊ TỚI HẠN PHÂN PHỐI FISHER

$$\boxed{P\left(F > f(n,m,\alpha)\right) = \alpha \ voi \ F \sim F(n;m)}$$

Bậc tự do n

m	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	4052.18	4999.50	5403.35	5624.58	5763.65	5858.99	5928.36	5981.07	6022.47	6055.85	6083.32	6106.32	6125.86	6142.67	6157.28
2	98.503	99.000	99.166	99.249	99.299	99.333	99.356	99.374	99.388	99.399	99.408	99.416	99.422	99.428	99.433
3	34.116	30.817	29.457	28.710	28.237	27.911	27.672	27.489	27.345	27.229	27.133	27.052	26.983	26.924	26.872
4	21.198	18.000	16.694	15.977	15.522	15.207	14.976	14.799	14.659	14.546	14.452	14.374	14.307	14.249	14.198
5	16.258	13.274	12.060	11.392	10.967	10.672	10.456	10.289	10.158	10.051	9.963	9.888	9.825	9.770	9.722
6	13.745	10.925	9.780	9.148	8.746	8.466	8.260	8.102	7.976	7.874	7.790	7.718	7.657	7.605	7.559
7	12.246	9.547	8.451	7.847	7.460	7.191	6.993	6.840	6.719	6.620	6.538	6.469	6.410	6.359	6.314
8	11.259	8.649	7.591	7.006	6.632	6.371	6.178	6.029	5.911	5.814	5.734	5.667	5.609	5.559	5.515
9	10.561	8.022	6.992	6.422	6.057	5.802	5.613	5.467	5.351	5.257	5.178	5.111	5.055	5.005	4.962
10	10.044	7.559	6.552	5.994	5.636	5.386	5.200	5.057	4.942	4.849	4.772	4.706	4.650	4.601	4.558
11	9.646	7.206	6.217	5.668	5.316	5.069	4.886	4.744	4.632	4.539	4.462	4.397	4.342	4.293	4.251
12	9.330	6.927	5.953	5.412	5.064	4.821	4.640	4.499	4.388	4.296	4.220	4.155	4.100	4.052	4.010
13	9.074	6.701	5.739	5.205	4.862	4.620	4.441	4.302	4.191	4.100	4.025	3.960	3.905	3.857	3.815
14	8.862	6.515	5.564	5.035	4.695	4.456	4.278	4.140	4.030	3.939	3.864	3.800	3.745	3.698	3.656
15	8.683	6.359	5.417	4.893	4.556	4.318	4.142	4.004	3.895	0,01	3.730	3.666	3.612	3.564	3.522
16	8.531	6.226	5.292	4.773	4.437	4.202	4.026	3.890	3.780	3.691	3.616	3.553	3.498	3.451	3.409
17	8.400	6.112	5.185	4.669	4.336	4.102	3.927	3.791	3.682	3.593	3.519	3.455	3.401	3.353	3.312
18	8.285	6.013	5.092	4.579	4.248	4.015	3.841	3.705	3.597	3.508	3.434	3.371	3.316	3.269	3.227
19	8.185	5.926	5.010	4.500	4.171	3.939	3.765	3.631	3.523	3.434	3.360	3.297	3.242	3.195	3.153
20	8.096	5.849	4.938	4.431	4.103	3.871	3.699	3.564	3.457	3.368	3.294	3.231	3.177	3.130	3.088