## Chương 1

# Một số khái niệm trong xác suất và thống kê mô tả

Một phần kiến thức cơ bản không thể tách rời trong quá trình thiết kế và xử lý dữ liệu thí nghiệm đó là các kiến thức về xác suất và thống kê. Mục đích của chương này là tập hợp lại một số khái niệm về xác suất, các phân phối thường được sử dụng trong sinh học nói chung và trong chăn nuôi, thú y nói riêng; đồng thời cũng khái quát hoá và nêu ý nghĩa của một số tham số thống kê mô tả cơ bản.

## 1.1. Tóm tắt về xác suất và biến ngẫu nhiên

### 1.1.1. Xác suất cơ bản

Số chỉnh hợp chập k trong n vật $A_n^k = n(n-1)(n-2)...(n-k+1) = \frac{n!}{(n-k)!}$

Số tổ hợp chập k của n vật $C_n^k = \frac{A_n^k}{k!} = \frac{n!}{k!(n-k)!}$

Số hoán vị của k vật $A_k^k = k!$

Số chỉnh hợp lặp chập k của n vật $\tilde{A}_n^k = n^k$

Nhị thức Niu-tơn $(a+b)^n = \sum_{k=0}^{n} C_n^k a^{n-k} b^k$

Quy tắc cộng tổng quát $p(A \cup B) = p(A) + p(B) - p(A \cap B)$

Quy tắc cộng đơn giản $p(A \cup B) = p(A) + p(B)$ nếu $A \cap B = \varnothing$

Quy tắc nhân tổng quát $p(A \cap B) = p(A).\ p(B/A) = p(B).p(A/B)$

Quy tắc nhân đơn giản $p(A \cap B) = p(A).\ p(B)$ nếu A, B độc lập

### 1.1.2. Hệ sự kiện đầy đủ

Hệ sự kiện đầy đủ hay hệ sự kiện toàn phần nếu:

$$\bigcup_{i=1}^{n} A_i = \Omega \quad \text{và} \quad A_i \cap A_j = \varnothing \text{ với } i \neq j$$

Công thức xác suất toàn phần $p(B) = \sum_{k=1}^{n} p(A_i).p(B/A_i)$

Công thức Bayes $p(A/B) = \dfrac{p(A_i).p(B/A_i)}{p(B)}$

### 1.1.3. Biến ngẫu nhiên, bảng phân phối, hàm phân phối

Kỳ vọng toán học $MX = \sum_{1}^{n} x_i p_i$

Phương sai $DX = \sum_{1}^{n} (x_i - MX)^2 p_i$ hay $DX = \sum_{i=1}^{n} x_i^2 p_i - (MX)^2$

Bảng phân phối của biến ngẫu nhiên rời rạc

| X | $x_1$ | $x_2$ | . . . | $x_n$ | Tổng |
|---|---|---|---|---|---|
| $p_i$ | $p_1$ | $p_2$ | . . . | $p_n$ | 1 |

Hàm phân phối

$$F(x) = p(X < x) = \begin{cases} 0 & x \le x_1 \\ p_1 & x_1 \le x < x_2 \\ p1 + p2 & x_2 \le x < x_3 \\ p1 + p2 + p3 & x_3 \le x < x_4 \\ \ldots & \\ 1 & x_n < x \end{cases}$$

### 1.1.4. Một số phân phối thường gặp

**Phân phối Bécnuli**

| X | 0 | 1 |
|---|---|---|
| $p_i$ | p | q |

Kỳ vọng MX = μ = p  Phương sai DX = pq

**Phân phối Nhị thức B(n,p)**

| X | 0 | 1 | . . . | K | . . . | n |
|---|---|---|---|---|---|---|
| $p_i$ | $q^n$ | $C^1_n pq^{n-1}$ | . . . | $C^k_n p^k q^{n-k}$ | . . . | $p^n$ |

MX = np  DX=npq
ModX là số nguyên
np-q ≤ ModX ≤np+p

**Phân phối siêu bội**

Nếu trong N bi có M bi trắng, rút n bi, X là số bi trắng

$$X = 0, n \text{ với } p_k = p(X = k) \frac{C_M^k C_{N-M}^{n-k}}{C_N^n}$$

$$MX = \frac{nM}{N} \qquad DX = n\frac{M}{N}\frac{N-M}{N}\frac{N-n}{N-1}$$

**Phân phối hình học**

$X = \overline{1, \infty}$ với $p_k = p(X = k) = pq^{k-1}$ (p là xác suất thành công, q = 1- p)

$$MX = \frac{1}{p} \qquad DX = \frac{q}{p^2}$$

**Phân phối Poátxông**

$X = \overline{0, \infty}$ với xác suất $p_k = p(X = k) = \frac{e^{-\lambda}}{k!}\lambda^k$

$$MX = DX = \lambda$$

**Phân phối chuẩn *N*(μ,σ2 )**

Hàm mật độ xác suất $f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$

$$p(a < X, b) = \Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})$$

với $\Phi(z)$ là hàm phân phối của biến chuẩn tắc

**Phân phối chuẩn tắc N(0,1)**

Mật độ xác suất $\varphi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$

Hàm phân phối $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} e^{-\frac{x^2}{2}} dx$

Tính gần đúng phân phối nhị thức bằng phân phối chuẩn khi n lớn

$$p(k \le X \le l) \approx \Phi(\frac{l-np}{\sqrt{npq}}) - \Phi(\frac{k-np}{\sqrt{npq}})$$

$$p(X = k) ) \approx \frac{1}{\sqrt{npq}} \varphi(\frac{k-np}{\sqrt{npq}})$$

Dung lượng mẫu cần thiết để trung bình cộng khác μ không quá ε (độ chính xác) khi có phân phối chuẩn $N(\mu,\sigma^2)$ và mức tin cậy P = 1 - α

$$n \ge \frac{z^2 \sigma^2}{\varepsilon^2}$$ z là giá trị sao cho Φ(z) = 1-α/2

Dung lượng mẫu cần thiết để tần suất khác xác suất không quá ε trong phân phối nhị thức và mức tin cậy P = 1 - α

$$n \ge \frac{z^2}{4\varepsilon^2}$$ z là giá trị sao cho Φ(z) = 1-α/2

## 1.2. Biến sinh học

Trong quá trình thực hiện thí nghiệm, chúng ta tiến hành thu thập dữ liệu để sau đó xử lý và đưa ra các kết luận. Các dữ liệu có thể là các giá trị bằng số hoặc bằng chữ đặc trưng cho một cá thể hoặc một nhóm và thay đổi từ cá thể này qua cá thể khác. Các dữ liệu như vậy được gọi là các biến, hay còn được gọi là các biến ngẫu nhiên vì các dữ liệu thu được là kết quả của việc chọn một cách ngẫu nhiên cá thể hay nhóm cá thể trong tổng thể.

### 1.2.1. Khái niệm về biến sinh học

Đối tượng nghiên cứu trong chăn nuôi là các vật sống, vì vậy các biến như đã nêu trên gọi chung là các biến sinh học. Có thể phân loại các biến sinh học như sau:

**Biến định tính (qualitative)**
Biến định danh (nominal)
Biến thứ hạng (ranked)

**Biến định lượng (quantitative)**
Biến liên tục (continuous)
Biến rời rạc (discontinuous)

**Biến định tính** bao gồm các **biến có hai trạng thái (binary)**: thí dụ như giới tính (cái hay đực), vật nuôi sau khi được điều trị (sống hay chết, khỏi bệnh hay không khỏi bệnh), tình trạng nhiễm bệnh (có, không), mang thai (có, không) . . .Tổng quát hơn có các **biến có nhiều trạng thái**, từ đó chia ra các lớp (loại) thí dụ mầu lông của các giống lợn (trắng, đen, loang, hung, . . .) các kiểu gen (đồng hợp tử trội, dị hợp tử, đồng hợp tử lặn . . . ); giống bò (bò vàng, Jersey, Holstein…). Các biến như thế được gọi là **biến định danh (nominal)** hay biến có thang đo định danh, cũng còn gọi là biến thuộc tính. Trong các biến có nhiều trạng thái, có một số biến có thể sắp thứ tự theo một cách nào đó, ví dụ mức độ mắc bệnh của vật nuôi. Thường dùng số thứ tự để xếp hạng các biến này, thí dụ xếp động vật theo mức độ mắc bệnh (--, -, -+, +, ++), thể trạng của vật nuôi (đối với bò từ 1-5, 1-rất gầy,…, 5-rất béo) . Các biến này gọi là **biến thứ hạng (ranked)** hay biến có thang đo thứ bậc.

**Biến định lượng** là biến phải dùng một gốc đo, một đơn vị đo để xác định giá trị (số đo) của biến. Biến định lượng bao gồm: **biến rời rạc**, thí dụ số trứng nở khi ấp 12 quả (X = 0, 1, . . . , 12), số lợn con sinh ra trong một lứa đẻ, số tế bào hồng cầu đếm trên đĩa của kính hiển vi và **biến liên tục**, thí dụ khối lượng gà 45 ngày tuổi, sản lượng sữa bò trong một chu kỳ, tăng trọng trên ngày của động vật, nồng độ canxi trong máu . . . Sau khi chọn đơn vị đo thì giá trị cụ thể của X là một số nằm trong một khoảng [a, b] nào đó.

Đối với các biến định lượng có thể phân biệt: 1) **biến khoảng** (interval) hay biến có thang đo khoảng, biến này chỉ chú ý đến mức chênh lệch giữa hai giá trị (giá trị 0 mang tính quy ước, tỷ số hai giá trị không có ý nghĩa). Thí dụ đối với nhiệt độ chỉ nói nhiệt độ tăng thêm hay giảm đi mấy °C ( thí dụ cơ thể đang từ 36,5°C tăng lên 38°C là biểu hiện bắt đầu sốt cao) chứ không nói vật thể có nhiệt độ 60°C nóng gấp đôi vật thể có nhiệt độ 30°C. Hướng gió có quy ước 0° là hướng Bắc, 45° là hướng Đông Bắc, 90° là hướng Đông, 180° là hướng Nam . . . , không thể nói hướng gió Đông gấp đôi hướng gió Đông Bắc; 2) **biến tỷ số** (ratio) hay biến có thang đo tỷ lệ, đối với biến này giá trị 0, mức chênh lệch giữa hai giá trị và tỷ số hai giá trị đều có ý nghĩa. Thí dụ khối lượng bắt đầu thí nghiệm của lợn là 25 kg, khối lượng kết thúc là 90 kg, vậy khối lượng kết thúc thí nghiệm nặng gấp 3,6 lần.

### 1.2.2. Tổng thể và mẫu

Một đám đông gồm rất nhiều cá thể chung nhau nguồn gốc, hoặc chung nhau nơi sinh sống, hoặc chung nhau nguồn lợi . . . được gọi là một tổng thể. Lấy từng cá thể ra đo một biến sinh học X, chúng ta được một biến ngẫu nhiên, có thể định tính hoặc định lượng. Tập hợp tất cả các giá trị của X gọi là một tổng thể (population).

Muốn hiểu biết đầy đủ về biến X phải khảo sát toàn bộ tổng thể, nhưng vì nhiều lý do không thể làm được. Có thể do không đủ tiền tài, vật lực, thời gian, . . . , nên không thể khảo sát toàn bộ, cũng có thể do phải huỷ hoại cá thể khi khảo sát nên không thể khảo sát toàn bộ, cũng có khi cân nhắc giữa mức chính xác thu được và chi phí khảo sát thấy không cần thiết phải khảo sát hết.

Như vậy là có nhiều lý do khiến người ta chỉ khảo sát một bộ phận gọi là mẫu (sample) sau đó xử lý các dữ liệu (số liệu) rồi đưa ra các kết luận chung cho tổng thể. Các kết luận này được gọi là "kết luận thống kê".

Để các kết luận đưa ra đúng cho tổng thể thì mẫu phải "phản ánh" được tổng thể (còn nói là mẫu phải "đại diện", phải "điển hình" cho tổng thể. . .), không được thiên về phía "tốt" hay thiên về phía "xấu".

### 1.2.3. Sơ lược về cách chọn mẫu

Tuỳ theo đặc thù của ngành nghề người ta đưa ra rất nhiều cách chọn mẫu khác nhau, thí dụ chọn ruộng để gặt nhằm đánh giá năng suất, chọn các sản phẩm của một máy để đánh giá chất lượng, chọn các hộ để điều tra dân số hoặc điều tra xã hội học, chọn một số sản phẩm ra kiểm tra trước khi xuất khẩu một lô hàng. . . Cách chọn mẫu phải hợp lý về mặt chuyên môn, phải dễ cho người thực hiện và phải đảm bảo yêu cầu chung về mặt xác suất thống kê là "ngẫu nhiên" không thiên lệch.

Thuần tuý về thống kê cũng có nhiều cách chọn mẫu:

Chọn mẫu hoàn toàn ngẫu nhiên (rút thăm, dùng bảng số ngẫu nhiên để lựa chọn,. . .).

Chia tổng thể thành các lớp đồng đều hơn theo một tiêu chuẩn nào đó thí dụ chia toàn quốc thành các vùng (vùng cao, trung du, đồng bằng), chia theo tầng lớp xã hội, chia theo thu nhập, theo ngành nghề, chia sản phẩm thành các lô hàng theo nguồn vật liệu, theo ngày sản xuất, . . . Sau khi có các lớp thì căn cứ vào mức đồng đều trong từng lớp mà chọn số lượng cá thể (dung lượng mẫu) đại diện cho lớp.

Có thể chia tổng thể thành các lớp, sau đó chọn một số lớp gọi là mẫu cấp một. Mỗi lớp trong mẫu cấp một lại được chia thành nhiều lớp nhỏ hơn, đều hơn. Chọn một số trong đó gọi là mẫu cấp hai. Có thể khảo sát hết các cá thể trong mẫu cấp hai hoặc chỉ khảo sát một bộ phận.

Không đi sâu vào việc chọn mẫu chúng ta chỉ nhấn mạnh mẫu phải ngẫu nhiên, phải chọn mẫu một cách khách quan không được chọn mẫu theo chủ quan người chọn.

### 1.2.4. Các tham số của mẫu

Gọi số cá thể được chọn vào mẫu là kích thước (cỡ, dung lượng) mẫu n. Gọi các số liệu đo được trên các cá thể của mẫu là $x_1, x_2, \ldots, x_n$, nếu có nhiều số liệu bằng nhau thì có thể ghi lại dưới dạng có tần số (số lần gặp)

| | | | | | |
|---|---|---|---|---|---|
| Giá trị $x_i$ | $x_1$ | $x_2$ | . . . | $x_k$ | |
| tần số $m_i$ | $m_1$ | $m_2$ | . . . | $m_k$ | $\sum_{i=1}^{k} m_i = n$ |

Các tham số (số đặc trưng) của mẫu, hay còn gọi là các thống kê, được chia thành hai nhóm: 1) các tham số về vị trí và 2) các tham số về độ phân tán của số liệu.

Các **tham số về vị trí** thường gồm: a) trung bình, b) trung vị, c) mode. Các **tham số về độ phân tán** gồm: a) phương sai, b) độ lệch chuẩn, c) sai số chuẩn, d) khoảng biến động và e) hệ số biến động.

## TRUNG BÌNH

**Trung bình cộng** ký hiệu là $\bar{x}$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \quad \text{hay} \quad \bar{x} = \frac{\sum_{i=1}^{k} x_i m_i}{\sum_{i=1}^{k} m_i} \text{ khi có tần suất}$$

**Ví dụ 1.1:** Khối lượng (gram) của 16 chuột cái tại thời điểm cai sữa như sau:

54,1 49,8 24,0 46,0 44,1 34,0 52,6 54,4
56,1 52,0 51,9 54,0 58,0 39,0 32,7 58,5

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{54{,}1 + 49{,}8 + .... + 58{,}5}{16} = \frac{761{,}2}{16} = 47{,}58 \text{ gram}$$

**Ví dụ 1.2:** Phân bố tần suất khối lượng của 4547 lợn Piétrain × (Yorkshire × Landrace) nuôi vỗ béo đến 210 ngày tuổi (kg).

| Nhóm khối lượng (kg) | Khối lượng trung bình (kg) | Số lượng | Tần suất | Tần suất tích luỹ |
|---|---|---|---|---|
| 60,73 - 66,99 | 63,86 | 11 | 0,24 | 0,24 |
| 67,00 - 74,99 | 71,00 | 31 | 0,68 | 0,92 |
| 75,00 - 82,99 | 79,00 | 80 | 1,76 | 2,68 |
| 83,00 - 90,99 | 87,00 | 218 | 4,79 | 7,48 |
| 91,00 - 98,99 | 95,00 | 484 | 10,64 | 18,12 |
| 99,00 - 106,99 | 103,00 | 951 | 20,91 | 39,04 |
| 107,00 - 114,99 | 111,00 | 1083 | 23,82 | 62,85 |
| 115,00 - 122,99 | 119,00 | 907 | 19,95 | 82,8 |
| 123,00 - 130,99 | 127,00 | 512 | 11,26 | 94,06 |
| 131,00 - 138,99 | 135,00 | 203 | 4,46 | 98,53 |
| 139,00 - 146,99 | 143,00 | 55 | 1,21 | 99,74 |
| 147,00 - 156,10 | 151,55 | 12 | 0,26 | 100,00 |

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i m_i}{\sum_{i=1}^{k} m_i} = \frac{63{,}86 \times 11 + 71{,}00 \times 31 + ..... + 151{,}55 \times 12}{11 + 31 + .... + 12} = 110{,}48\,\text{kg}$$

Giá trị trung bình cộng có bất lợi là bị các giá trị ngoại lai làm ảnh hưởng. Giá trị ngoại lai là giá trị có xu hướng không thích hợp với toàn bộ số liệu thu thập được, thường là các giá trị quá lớn hoặc quá bé so với bình thường. Nếu giá trị ngoại lai quá lớn sẽ làm cho giá trị trung bình có xu hướng tăng quá mức hoặc ngược lại.

**Trung bình nhân** ký hiệu là G

$$G = \sqrt[n]{x_1 x_2 ... x_n} \qquad G = \sqrt[n]{x_1^{m_1} x_2^{m_2} ... x_k^{m_k}}$$

**Ví dụ 1.3:** Bệnh dại đã tăng 10% trong năm thứ nhất, 11% trong năm thứ 2 và 15% trong năm thứ 3. Mức tăng trưởng trung bình của bệnh là bao nhiêu phần trăm?

Ta không thể tính tăng trưởng trung bình như sau (10 + 11 + 15)/3 = 12 mà phải tính mức tăng trưởng trung bình là $G = \sqrt[n]{x_1 x_2 ... x_n} = \sqrt[3]{1{,}1 \times 1{,}11 \times 1{,}15} = 1{,}11979$. Nghĩa là mức tăng trưởng trung bình là 0,11979 hay tương đương mức 11,979 %.

**Ví dụ 1.4:** Một loại mô bào sinh trưởng sau 3 tháng sẽ tăng gấp đôi khối lượng. Mức tăng trưởng trung bình mỗi tháng là bao nhiêu?

Mức tăng trưởng trung bình mỗi tháng là: $G = \sqrt[3]{2}$ = 1,26; nghĩa là 26% mỗi tháng.
Ta có thể minh hoạ sự tăng trưởng qua 3 tháng như sau:

1×1,26 = 1,26

1,26×1,26 = 1,5876

1,5876×1,26 = 2,00037

**Trung bình điều hoà** ký hiệu là H

$$H = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} \qquad \text{hoặc} \qquad H = \frac{n}{\sum_{i} \frac{m_i}{x_i}}$$

**Ví dụ 1.5:** Ba lò mổ mỗi lò mổ 1000 con; lò mổ thứ nhất có năng suất giết mổ 10 con/giờ, lò mổ thứ hai 15 con/giờ và lò mổ thứ ba 30 con/giờ. Trung bình một giờ giết mổ được bao nhiêu con?

Trung bình sẽ không phải là (10 + 15 + 30)/3 = 55/3. Đây là trung bình cộng, chính bằng trung bình mỗi giờ nếu cả 3 lò mổ song song song với nhau.

Giá trị trung bình phải là $H = \frac{n}{\sum_{i} \frac{1}{x_i}} = \frac{3}{\frac{1}{10} + \frac{1}{15} + \frac{1}{30}} = 15$ con/giờ.

Điều này có thể minh hoạ như sau: Để giết mổ được 90 con lò thứ nhất phải thực hiện trong 9 giờ, lò thứ hai trong 6 giờ và lò thứ 3 trong 3 giờ; nghĩa là 270 con lợn được giết mổ trong 18 giờ; tức là trung bình 15 con/giờ. Chú ý rằng số lợn giết mổ được cố định khi bắt đầu.

**TRUNG VỊ** ký hiệu Me

Nếu sắp xếp các giá trị từ nhỏ đến lớn thì giá trị ở vị trí chính giữa được gọi là trung vị (Me). Nói một cách lý thuyết thì Me là giá trị có 50% số giá trị nhỏ hơn và 50% số giá trị lớn hơn. Để tính nhanh giá trị trung vị ta có thể tiến hành các bước sau:

1) Sắp xếp các giá trị theo trình tự tăng dần
2) Đánh số thứ tự cho các dữ liệu
3) Tìm trung vị ở vị trí có số thứ tự (n + 1)/2

Nếu n là số lẻ và các giá trị đều khác nhau thì có một giá trị chính ở giữa

**Ví dụ 1.6:** Nồng độ vitamin E (µmol/l) của 11 bê cái có dấu hiệu lâm sàng của phát triển cơ không bình thường được trình bày như sau:

4,2 3,3 7,0 6,9 5,1 3,4 2,5 8,6 3,5 2,9 4,9

Sau khi sắp xếp theo thứ tự tăng dần ta có:

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| 2,5 | 2,9 | 3,3 | 3,4 | 3,5 | **4,2** | 4,9 | 5,1 | 6,9 | 7,0 | 8,6 |
| 1 | 2 | 3 | 4 | 5 | **6** | 7 | 8 | 9 | 10 | 11 |

Như vậy vị trí trung vị sẽ là (n + 1)/2 = (11 + 1)/2 = 6, do 6 là vị trí của trung vị nên giá trị của trung vị sẽ là 4,2.

Nếu n là số chẵn và các giá trị đều khác nhau thì có 2 số đứng giữa, cả hai đều được gọi là trung vị. Khoảng giữa 2 số đứng giữa được gọi là khoảng trung vị. Nếu được phép dùng số thập phân thì lấy điểm giữa của khoảng làm trung vị Me.

**Xét ví dụ 1.1:** Khối lượng (gram) của 16 chuột cái tại thời điểm cai sữa như sau:

54.1 49.8 24.0 46.0 44.1 34.0 52.6 54.4

56.1 52.0 51.9 54.0 58.0 39.0 32.7 58.5

Vị trí của trung vị sẽ là (16 + 1)/2 = 8,5; khoảng trung vị sẽ nằm ở vị trí số 8 và số 9, tức là từ 49,8 – 51,9. Như vậy giá trị của trung vị Me = (49,8 + 51,9)/2 = 50,9.

Nếu các số liệu chia thành lớp có tần số thì phải chọn lớp trung vị sau đó nội suy để tính gần đúng trung vị.

Ngoài trung vị còn có các phân vị, trong đó hay dùng nhất là tứ phân vị dưới $Q_1$ mà chúng ta có thể định nghĩa một cách lý thuyết là giá trị có 25% số giá trị nhỏ hơn, tứ phân vị trên $Q_2$ là giá trị có 25% số giá trị lớn hơn.

**MODE** ký hiệu Mod

Mode là giá trị có tần suất cao nhất. Thông thường Mode có giá trị khác với giá trị trung bình cộng và trung vị. Ba giá trị này này sẽ bằng nhau khi số liệu có phân bố chuẩn. Nhóm Mode hay lớp Mode là nhóm hoặc lớp mà một số lớn các quan sát rơi vào đó. Thông qua tổ chức đồ ta có thể xác định được giá trị của lớp này.

Xét trường hợp ví dụ 2, nhóm Mod được đại diện bằng các giá trị từ 107 đến 115 kg. Từ 4547 lợn quan sát có 1083 con nằm trong khoảng từ 107 đến 115kg ; đây là tần suất cao nhất. Cũng theo ví dụ 1 ta thấy Mod có giá trị khoảng 111kg.

| P | 60,7 | 67,0 | 75,0 | 83,0 | 91,0 | 99,0 | **107,0** | 115,0 | 123,0 | 131,0 | 139,0 | 147,0 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| (kg) | 66,9 | 74,9 | 82,9 | 90,9 | 98,9 | 106,9 | **114,9** | 122,9 | 130,9 | 138,9 | 146,9 | 156,1 |
| n | 11 | 31 | 80 | 218 | 484 | 951 | **1083** | 907 | 512 | 203 | 55 | 12 |

Trường hợp có nhiều giá trị có tần số lớn bằng nhau và lớn hơn các tần số khác thì không xác định được Mod.

Trường hợp số liệu chia lớp thì tìm lớp có tần số lớn nhất sau đó dùng cách nội suy để tính gần đúng Mod.

**PHƯƠNG SAI MẪU** ký hiệu **s²**

Phương sai mẫu chưa hiệu chỉnh $s^2_p$ tính theo công thức:

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

Phương sai mẫu được dùng trong tài liệu này là **phương sai đã hiệu chỉnh**, gọi tắt là phương sai mẫu $s^2$:

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

Đối với máy tính bỏ túi, có thể tính phương sai theo công thức sau:

$$s^2 = \frac{(\sum_i x_i^2 - \frac{(\sum_i x_i)^2}{n})}{(n-1)}$$

Khi có phương sai mẫu chưa hiệu chỉnh $s^2_p$ có thể tính $s^2$ theo công thức

$$s^2 = \frac{n}{(n-1)} s_p^2$$

**Xét ví dụ 1.1**, khối lượng của 16 chuột cái tại thời điểm cai sữa; giá trị trung bình đã tính là 47,58gram. Như vậy phương sai mẫu hiệu chỉnh sẽ là:

$$s^2 = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}{n-1} = \frac{(54{,}1-47{,}58)^2 + (49{,}8-47{,}58)^2 + .... + (58{,}5-47{,}58)^2}{16-1} = 103{,}27 \text{ gram}^2$$

**ĐỘ LỆCH CHUẨN** ký hiệu là **s**

Căn bậc hai của $s^2$ gọi là độ lệch chuẩn: $s = \sqrt{s^2}$

Xét ví dụ 1, khối lượng của 16 chuột cái tại thời điểm cai sữa. Các số liệu này đã được sử dụng để tính giá trị trung bình (47,58 gram) và phương sai (103,27 gram²) như đã nêu trên.

Như vậy độ lệch chuẩn sẽ là: $s = \sqrt{s^2} = \sqrt{103{,}27} = 10{,}16 \text{ gram}$

**HỆ SỐ BIẾN ĐỘNG** ký hiệu là **Cv (%)**

Hệ số biến động được tính theo công thức

$$Cv = \frac{s}{\bar{x}} \times 100$$

**Xét ví dụ 1.1**, khối lượng của 16 chuột cái tại thời điểm cai sữa. Ta đã có giá trị trung bình (47,58gram) và độ lệch chuẩn (10,16 gram). Như vậy phương sai mẫu hiệu chỉnh sẽ là:

$$Cv = \frac{s}{\bar{x}} \times 100 = \frac{10,16}{47,58} \times 100 = 21,36\,\%$$

**KHOẢNG BIẾN THIÊN** (**phạm vi chứa số liệu Range**)

Gọi $X_{max}$ là giá trị lớn nhất, Gọi $X_{min}$ là giá trị nhỏ nhất, ta có khoảng biến thiên:

$$R = x_{max} - x_{min}$$

**Với ví dụ 1.1**, khối lượng của 16 chuột tại thời điểm cai sữa.

Ta có $R = x_{max} - x_{min} = 58,5 – 24,0 = 34,5$ gram

**SAI SỐ CHUẨN** (**sai số của trung bình cộng**) ký hiệu là SE

$$SE = \frac{S}{\sqrt{n}}$$

**Xét ví dụ 1.1**, khối lượng của 16 chuột cái tại thời điểm cai sữa. Ta đã có độ lệch chuẩn (10,16 gram). Như vậy sai số tiêu chuẩn sẽ là:

$$SE = \frac{S}{\sqrt{n}} = \frac{10,16}{\sqrt{16}} = 2,54\,\text{gram}$$

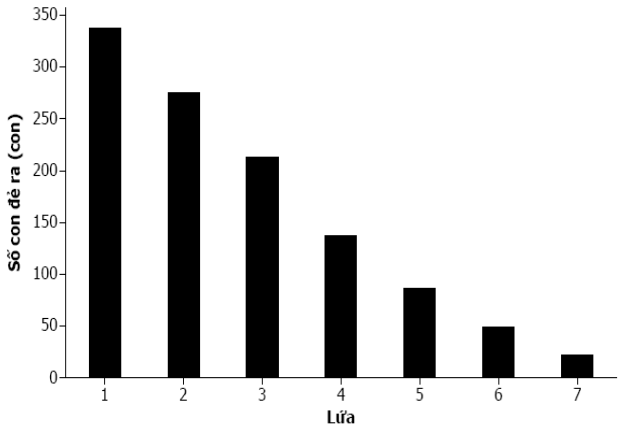
Ngoài các tham số trên, trong thống kê còn dùng độ lệch (độ bất đối xứng), độ nhọn. Hai tham số này được dùng khi xem xét có nên chuyển đổi số liệu không phân phối chuẩn thành số liệu phân phối chuẩn hay không.

### 1.2.5. Biểu diễn số liệu bằng đồ thị

Đồ thị là tóm tắt số liệu ở các dạng hình ảnh khác nhau và cho phép dễ dàng phát hiện những điểm đặc biệt hơn so với tóm tắt bằng số. Đồ thị đặc biệt hiệu quả khi ta muốn biết được các thông tin về số liệu một cách nhanh chóng.

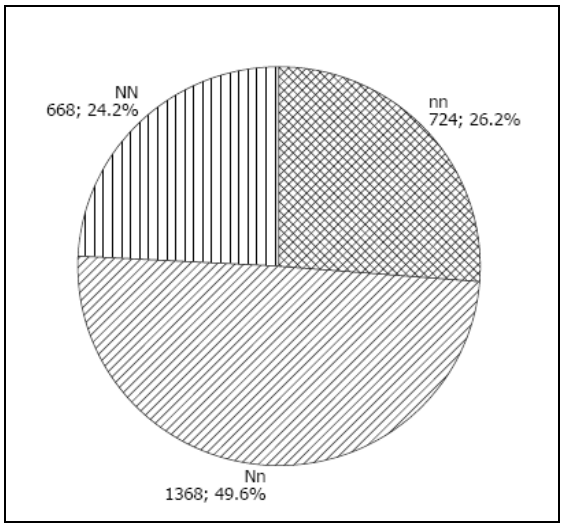
Có nhiều cách biểu diễn số liệu bằng đồ thị: Đồ thị tần số, đồ thị hình thanh, đồ thị đa giác, chữ nhật (tổ chức đồ).

Đối với biến định tính hoặc biến rời rạc có thể biểu diễn số liệu bằng đồ thị thanh hoặc đồ thị bánh hình tròn.




| Lứa | Số con đẻ ra (con) | Tần suất (%) | Tần suất tích luỹ (%) |
|---|---|---|---|
| 1 | 337 | 30,12 | 30,12 |
| 2 | 275 | 24,58 | 54,69 |
| 3 | 213 | 19,03 | 73,73 |
| 4 | 137 | 12,24 | 85,97 |
| 5 | 86 | 7,69 | 93,66 |
| 6 | 49 | 4,38 | 98,03 |
| 7 | 22 | 1,97 | 100,00 |

**Biểu đồ hình thanh** biểu diễn số lợn sơ sinh qua 7 lứa (n = 1119)

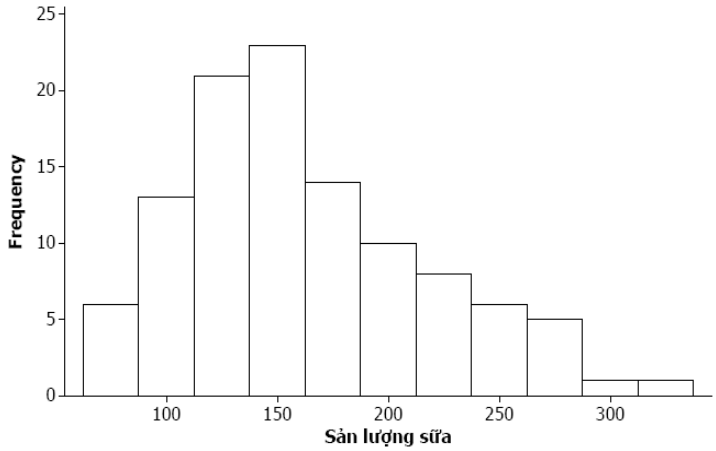



**Biểu đồ dạng bánh** biểu hiện tần số kiểu gen Halothane của lợn sơ sinh Pietrain (n =2760)

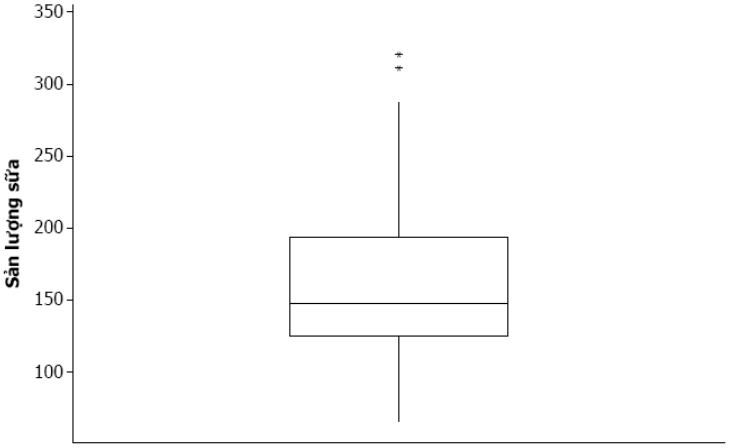
| Kiểu gen | Số con đẻ ra (con) | Tần suất (%) |
|---|---|---|
| nn | 724 | 26,20 |
| Nn | 1368 | 49,60 |
| NN | 668 | 24,20 |

Đối với biến định lượng có thể sử dụng đồ thị đa giác, đồ thị hộp hay tổ chức đồ để thể hiện. Ví dụ : Sản lượng sữa (kg) của 108 dê Bách Thảo trong một chu kỳ tiết sữa ghi lại như sau :

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 147,9 | 125,4 | 104,1 | 164,4 | 193,8 | 188,4 | 222,4 | 287,3 | 158,1 |
| 132,0 | 224,0 | 163,8 | 153,3 | 100,6 | 219,5 | 130,4 | 114,0 | 182,1 |
| 156,9 | 66,3 | 140,6 | 128,3 | 193,2 | 127,1 | 125,0 | 129,9 | 89,7 |
| 254,4 | 240,3 | 148,2 | 190,0 | 176,7 | 73,8 | 147,9 | 222,7 | 191,6 |
| 174,3 | 211,0 | 214,5 | 169,5 | 115,0 | 193,6 | 168,0 | 196,9 | 87,3 |
| 144,4 | 138,4 | 171,6 | 100,0 | 125,6 | 283,9 | 116,5 | 71,0 | 220,1 |
| 139,7 | 140,7 | 270,5 | 176,8 | 155,0 | 163,5 | 161,6 | 152,0 | 141,0 |
| 180,0 | 202,6 | 112,8 | 153,5 | 77,9 | 140,7 | 136,4 | 272,3 | 90,0 |
| 197,5 | 96,8 | 96,8 | 137,8 | 150,4 | 101,5 | 132,0 | 146,3 | 242,3 |
| 311,0 | 118,7 | 146,6 | 184,2 | 243,8 | 260,7 | 279,2 | 135,9 | 109,5 |
| 96,8 | 119,0 | 109,3 | 143,8 | 102,9 | 229,3 | 244,2 | 137,1 | 143,6 |
| 130,6 | 72,0 | 105,1 | 135,0 | 320,4 | 182,2 | 217,8 | 172,5 | 136,4 |



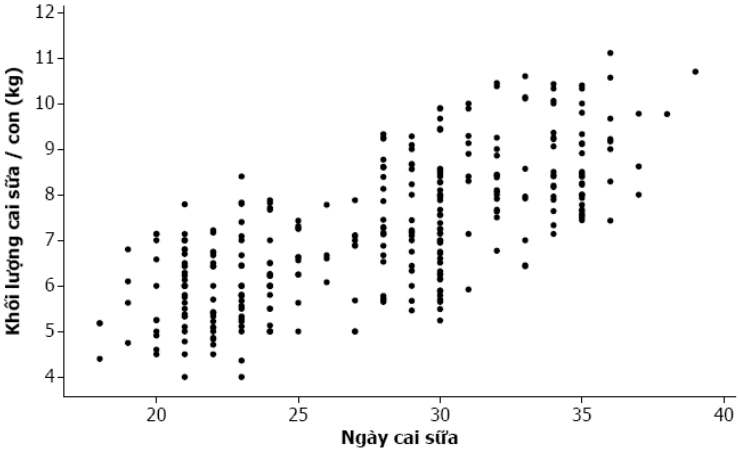

**Tổ chức đồ :** Phân bố tần suất sản lượng sữa dê Bách Thảo trong chu kỳ tiết sữa




**Đồ thị hộp :** Phân bố tần suất sản lượng sữa dê Bách Thảo trong chu kỳ tiết sữa

Tóm tắt và biểu diễn dữ liệu của các tính trạng số lượng (dữ liệu 2 chiều)

Đồ thị phân tán được sử dụng một cách rất hữu hiệu khi ta quan tâm đến mối liên hệ giữa 2 biến liên tục. Đồ thị được xây dựng khi ta vẽ n các điểm trên hệ toạ độ, các điểm này có toạ độ là $x_i y_i$. Vấn đề này sẽ được đề cập cụ thể trong chương 6.




**Đồ thị phân tán** thể hiện mối quan hệ giữa thời gian cai sữa (ngày) và khối lượng sơ sinh sinh/con (kg) của lợn Landrace n = 321.

## 1.3. Bài tập

### 1.3.1

Xác suất mắc một bệnh là P = 0,35 (0,35 là xác suất nhiễm bệnh được tính toán dựa trên một quan sát với dung lượng mẫu lớn). Hãy tính xác suất mắc bệnh của 2 trong số 10 động vật.

### 1.3.2

Xác suất mắc một bệnh là 0,25. Hãy tính xác suất không phát hiện được ca nhiễm bệnh trong số 30 động vật kiểm tra.

### 1.3.3

Bệnh dại xuất hiện với tần suất 0,005. Cần tiến hành kiểm tra bao nhiêu chó trong vùng để phát hiện bệnh dại với độ chính xác 95%.

### 1.3.4

Khối lượng (kg) ở 210 ngày tuổi của lợn Pietrain có các kiểu gen Halothane khác nhau được trình bày ở bảng số liệu dưới đây. Vẽ đồ thị và tính các tham số thống kê mô tả của bộ số liệu vừa nêu.

| NN | | | | Nn | | | | Nn | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 118,54 | 133,90 | 105,85 | 102,00 | 112,77 | 115,42 | 109,76 | 115,66 | 107,23 | 109,76 | 101,20 | 96,39 |
| 123,66 | 127,07 | 100,49 | 109,76 | 82,20 | 109,76 | 93,73 | 98,07 | 109,16 | 91,81 | 104,58 | 112,29 |
| 97,10 | 136,34 | 108,54 | 110,73 | 108,78 | 102,00 | 129,27 | 100,00 | 102,89 | 115,90 | 111,81 | 106,27 |
| 96,30 | 120,10 | 80,00 | 123,90 | 105,78 | 101,69 | 81,20 | 120,98 | 99,02 | 107,23 | 107,71 | 134,63 |
| 112,20 | 107,60 | 106,27 | 110,70 | 117,07 | 115,12 | 100,96 | 118,05 | 114,94 | 86,02 | 104,34 | 108,92 |
| 124,40 | 102,68 | 121,95 | 117,60 | 105,78 | 109,00 | 109,02 | 111,00 | 101,93 | 93,01 | 86,51 | 130,98 |
| 109,51 | 89,50 | 111,50 | 135,37 | 101,46 | 100,98 | 113,25 | 125,06 | 110,84 | 95,85 | 94,70 | 114,94 |
| 110,98 | 119,02 | 130,00 | 78,29 | 98,50 | 111,71 | 102,93 | 145,37 | 88,43 | 104,58 | 114,70 | 98,05 |
| 128,80 | 125,61 | 112,20 | 95,00 | 107,95 | 107,80 | 112,29 | 125,54 | 97,32 | 130,60 | 108,19 | 90,36 |
| 119,51 | 94,70 | 110,49 | 102,17 | 118,00 | 118,78 | 121,69 | 120,24 | 113,98 | 113,17 | 99,27 | 123,13 |
| 120,24 | 91,33 | 101,20 | 103,61 | 96,39 | 91,22 | 126,83 | 116,63 | 117,83 | 104,34 | 131,08 | 111,57 |
| 114,10 | 114,60 | 137,56 | 92,44 | 121,95 | 92,00 | 104,34 | 89,76 | 120,24 | 90,36 | 102,65 | 91,71 |
| 100,20 | 144,88 | 122,68 | 116,30 | 114,22 | 97,59 | 107,00 | 111,57 | 107,56 | 88,67 | 106,34 | 105,78 |
| 114,00 | 102,89 | 102,00 | 113,66 | 111,81 | 99,76 | 124,39 | 105,12 | 129,76 | 108,43 | 95,85 | 104,82 |
| 104,15 | 116,80 | 116,34 | 67,07 | 105,78 | 118,05 | 120,96 | 121,95 | 119,76 | 113,90 | 115,37 | 114,39 |
| 101,71 | 117,56 | 116,63 | 119,28 | 111,33 | 95,66 | 95,85 | 99,27 | 110,49 | 105,54 | 104,10 | 110,36 |
| 86,27 | 112,44 | 111,22 | 102,41 | 113,73 | 101,70 | 96,10 | 109,27 | 110,36 | 133,01 | 118,54 | 109,40 |
| 106,34 | 116,34 | 111,50 | 126,59 | 97,56 | 108,67 | 110,36 | 103,13 | 110,73 | 111,95 | 97,56 | 104,10 |
| 110,49 | 117,11 | 112,00 | 108,78 | 100,00 | 105,61 | 131,95 | 122,65 | 81,93 | 65,85 | 111,33 | 102,17 |
| 128,54 | 136,10 | 121,71 | 131,71 | 125,61 | 74,88 | 108,00 | 96,87 | 101,93 | 118,78 | 120,96 | 120,98 |
| 112,68 | 111,57 | 103,66 | 96,34 | 121,93 | 118,00 | 126,99 | 93,66 | 105,54 | 97,11 | 94,94 | 126,10 |
| 107,47 | 120,00 | 131,95 | 88,29 | 101,46 | 107,95 | 84,10 | 85,37 | 93,90 | 123,37 | 81,22 | 108,43 |
| 103,90 | 110,98 | 104,15 | 74,15 | 108,92 | 112,53 | 105,61 | 111,08 | 95,18 | 111,33 | 111,33 | 96,59 |
| 101,50 | 113,20 | 121,50 | 121,50 | 91,00 | 138,07 | 92,68 | 94,15 | 105,78 | 122,20 | 109,40 | 116,63 |
| 114,88 | 83,90 | 153,70 | 120,50 | 103,00 | 108,54 | 76,39 | 106,75 | 93,01 | 96,63 | 110,60 | 109,88 |