

**ỦY BAN NHÂN DÂN TỈNH ĐỒNG THÁP
TRƯỜNG CAO ĐẲNG NGHỀ ĐỒNG THÁP**



**GIÁO TRÌNH
MÔN HỌC: THỐNG KÊ SINH HỌC
NGÀNH, NGHỀ: THÚ Y
TRÌNH ĐỘ: TRUNG CẤP/CAO ĐẲNG**

(Ban hành kèm theo Quyết định số 257/QĐ-TCĐNĐT-ĐT ngày 13 tháng 07 năm 2017
của Hiệu trưởng trường Cao đẳng Nghề Đồng Tháp)

Đồng Tháp, năm 2017

TUYÊN BỐ BẢN QUYỀN

Đây là giáo trình nội bộ của Trường Cao đẳng nghề Đồng Tháp nên các nguồn thông tin có thể được phép dùng nguyên bản hoặc trích dùng cho các mục đích về đào tạo và tham khảo.

Mọi mục đích lệch lạc hoặc sử dụng với ý đồ kinh doanh thiếu lành mạnh sẽ bị nghiêm cấm.

Tailieu.vn

LỜI NÓI ĐẦU

Thống kê sinh học, thuật ngữ này có thể được hiểu theo hai nghĩa: Thứ nhất: thống kê là các số liệu được thu thập để phản ánh các hiện tượng kinh tế - xã hội, tự nhiên, kỹ thuật. Thứ hai: thống kê là hệ thống các phương pháp được sử dụng để mô tả các hiện tượng kinh tế - xã hội, tự nhiên, kỹ thuật. Nói một cách tổng quát thống kê là hệ thống các phương pháp dùng để thu thập, xử lý và phân tích các con số (mặt lượng) của các hiện tượng để tìm hiểu bản chất và tính qui luật vốn có của chúng (mặt chất) trong điều kiện thời gian và không gian cụ thể. Thống kê thường được chia thành hai lãnh vực: - Thống kê mô tả(Descriptive Statistics): bao gồm các phương pháp thu thập số liệu, mô tả và trình bày số liệu, tính toán các đặc trưng đo lường. - Thống kê suy diễn(Inferential Statistics): bao gồm các phương pháp như ước lượng, kiểm định, phân tích mối liên hệ, dự đoán. trên cơ sở các thông tin thu thập từ mẫu.

Một số thuật ngữ dùng trong bố trí thí nghiệm n đơn vị thí nghiệm (experimental unit): vật liệu à tác động một hoặc một số nhân tố là đo lường các ảnh hưởng của nó. nhân tố (factor) là nguyên nhân gây ảnh hưởng đến các giá trị quan sát là bao gồm các mức độ khác nhau. nghiệm thức (treatment) có thể bao gồm các mức độ khác nhau của một nhân tố hoặc một phối hợp các mức độ của các nhân tố khác nhau mà ta muốn khảo sát ảnh hưởng của nó trên vật liệu thí nghiệm. sai số thí nghiệm (experimental error) là tổng cộng các nguồn biến động không kiểm soát được. Nguồn biến động luôn hiện hữu trong vật liệu thí nghiệm do phương pháp thực hiện thí nghiệm hoặc do người làm thí nghiệm.

Chắc chắn rằng, lần soạn thảo đầu tiên và chúng tôi cũng quan niệm rằng không một giáo trình hoặc bài giảng nào là hoàn hảo cả. Vì vậy, chúng tôi rất mong sự đóng góp ý kiến của quý vị để lần tái bản sau đạt yêu cầu cao hơn.

Chúng tôi rất chân thành cảm ơn!

CHƯƠNG TRÌNH MÔN HỌC

Tên môn học: Thống kê sinh học

Mã môn học: MH 22

Thời gian môn học: 30 giờ (Lý thuyết: 14 giờ ; Thực hành: 16 giờ)

I. Vị trí, tính chất của môn học

- Vị trí của môn học: Môn học được bố trí ở vị trí số 23 trong chương trình đào tạo cao đẳng nghề thú y.
- Tính chất của môn học: Là môn học cơ sở trong chương trình đào tạo cao đẳng nghề thú y. Giúp cho các em xử lý số liệu dự báo được một số biến cố có thể xảy ra.

II. Mục tiêu của môn học

- Nắm vững được các phương pháp, thuật toán trong việc tổng hợp, phân tích và luận giải những dữ kiện.
- Áp dụng được các thuật toán để chứng minh tính đúng đắn của các dữ kiện, các biến cố.
- Xử lý được các số liệu sinh học thu thập qua các thí nghiệm, nghiên cứu khoa học bằng phương pháp thống kê.
- Dự đoán, dự báo được một số biến cố, sự kiện sau khi đã có số liệu nghiên cứu khoa học và đã xử lý bằng phương pháp thống kê sinh học.
- Nghiêm túc, cẩn thận khi tư duy và xử lý số liệu thống kê.

III. Nội dung môn học

CHƯƠNG 1. GIỚI THIỆU VỀ THỐNG KÊ SINH HỌC

I. MỘT SỐ KHÁI NIỆM

1. Thống kê (Statistics)

Thuật ngữ này có thể được hiểu theo hai nghĩa:

Thứ nhất: thống kê là các số liệu được thu thập để phản ánh các hiện tượng kinh tế - xã hội, tự nhiên, kỹ thuật.

Thứ hai: thống kê là hệ thống các phương pháp được sử dụng để mô tả các hiện tượng kinh tế - xã hội, tự nhiên, kỹ thuật.

Nói một cách tổng quát thống kê là hệ thống các phương pháp dùng để thu thập, xử lý và phân tích các con số (mặt lượng) của các hiện tượng để tìm hiểu bản chất và tính qui luật vốn có của chúng (mặt chất) trong điều kiện thời gian và không gian cụ thể.

Thống kê thường được chia thành hai lãnh vực:

- *Thống kê mô tả* (Descriptive Statistics): bao gồm các phương pháp thu thập số liệu, mô tả và trình bày số liệu, tính toán các đặc trưng đo lường.
- *Thống kê suy diễn* (Inferential Statistics): bao gồm các phương pháp như ước lượng, kiểm định, phân tích mối liên hệ, dự đoán... trên cơ sở các thông tin thu thập từ mẫu.

2. Thống kê sinh học (Biometry)

Theo nghĩa hẹp, **biometry** bắt nguồn từ tiếng Hy Lạp *bios* = sự sống và *metron*

= đo đạc nên có người gọi đây là sinh trắc (biological measurement). Theo nghĩa rộng thì thống kê sinh học là khoa học về sự ứng dụng các phương pháp thống kê để giải quyết các vấn đề của sinh học vì vậy biometry cũng còn được gọi là *biological statistics* hoặc đơn giản là *biostatistics*. Các phương pháp thống kê bao gồm các bước (1) bố trí thí nghiệm, (2) thu thập dữ liệu, (3) trình bày và tóm tắt dữ liệu, (4) từ các dữ liệu mẫu suy rộng ra tổng thể.

Trong giáo trình này chúng ta sẽ tập trung vào hai lãnh vực có liên quan mật thiết với thống kê sinh học. Đó là bố trí thí nghiệm (experimental design) và phân tích thống kê (statistical analysis).

3. Mẫu và Tổng thể

Trong thống kê sinh học các nghiên cứu thường dựa trên quan sát riêng rẽ (individual observation), là những quan sát hoặc đo đạc tiến hành trên đơn vị mẫu nhỏ nhất (smallest sampling unit). Trong sinh học, đơn vị mẫu nhỏ nhất thường là cá thể. Nếu ta đo trọng lượng của 100 con chuột thì trọng lượng của mỗi con chuột chính là một quan sát. 100 trọng lượng chuột đo được đại diện cho một mẫu của quan sát (sample of observations). Nếu chúng ta nghiên cứu sự thay đổi trọng lượng của một con chuột trong một thời kỳ xác định thì mẫu sẽ là tất cả các số đo trọng lượng của con chuột đó trong suốt thời kỳ nghiên cứu.

Tuy nhiên, thuật ngữ “quan sát riêng rẽ” và “mẫu của quan sát” chỉ nên được cấu trúc mà không nêu được bản chất của dữ liệu nghiên cứu. Giá trị thật sự đo được thật sự bởi một quan sát riêng rẽ là một biến số (variable). Trong một đơn vị mẫu nhỏ nhất có thể có nhiều hơn một biến.

Tập hợp tất cả các dữ liệu mà ta quan tâm nghiên cứu trong một lãnh vực nào đó được gọi là tổng thể (population) hay còn gọi là tập hợp chính. Nếu ta chọn 5 người để nghiên cứu số lượng bạch cầu trong máu họ và từ đó rút ra kết luận về số lượng

bạch cầu trong máu của toàn bộ loài người thì tổng thể chính là toàn bộ loài người. Thông thường kích thước của tổng thể (N) rất lớn, thậm chí là vô hạn. Tổng thể có thể được mô tả bởi các thông số (parameters) nhưng thường các giá trị này không thể xác định chính xác được.

II. DỮ LIỆU TRONG SINH HỌC (DATA IN BIOLOGY)

1. Dữ liệu (Data)

Các ghi nhận, mô tả hoặc các thuộc tính, sự kiện, các quá trình đều có thể hình thành một khối dữ liệu. Các dữ liệu thường được đo ở dạng thang số (numerical scale) hoặc phân loại thành nhóm (category) rồi sau đó mã hoá dưới dạng số. Dưới đây là một số thí dụ về dữ liệu:

(1) Huyết áp tối thiểu của tất cả các học sinh ở một trường trung học được đo để xác định xem có bao nhiêu phần trăm học sinh có huyết áp tối thiểu trên 90 mm Hg. Trong trường hợp này dữ liệu là số đo huyết áp.

(2) Tất cả cán bộ viên chức của một công ty được yêu cầu báo cáo thể trọng hàng tháng để đánh giá hiệu quả của chương trình kiểm soát thể trọng. Dữ liệu là?

(3) Trong một nghiên cứu về giáo dục, tất cả các tài xế của một trường đại học được yêu cầu trả lời câu hỏi “bạn có lái xe khi cơ thể đã có rượu, bia hay không?”. Dữ liệu là gì?

Việc thu thập đủ và đúng dữ liệu rất quan trọng. Bạn không thể có một nghiên cứu tốt nếu không có các dữ liệu tốt.

2. Các quan sát (Observations) và các biến số (Variables)

Trong thống kê, các đối tượng nghiên cứu được gọi là các đơn vị quan sát (observational units). Trên đơn vị này, các đặc tính được quan sát hoặc đo đạc được gọi là các biến số (variables). Trong mỗi đối tượng nghiên cứu, các giá trị số gán cho biến số được gọi là các quan sát (observations) hay các biến (variate).

Thí dụ: để nghiên cứu huyết áp của các sinh viên trong một trường đại học, các nhà nghiên cứu đo huyết áp tối đa và tối thiểu cho từng sinh viên. Huyết áp tối đa và tối thiểu là các biến số, số đo huyết áp là các quan sát, các sinh viên là các đơn vị quan sát.

Trong mỗi đơn vị, chúng ta thường quan sát nhiều hơn một biến. Chẳng hạn, trong các sinh viên đã nêu trên, người ta nghiên cứu bệnh cao huyết áp ở 500 người. Ngoài các số đo huyết áp tối đa và tối thiểu, nhà nghiên cứu còn ghi nhận tuổi, chiều cao, giới tính, trọng lượng cơ thể. Trong trường hợp này, chúng ta có một bộ dữ liệu của 500 sinh viên với các quan sát được ghi nhận cho mỗi biến trong sáu biến của từng đơn vị quan sát.

3. Các loại thang đo (scale)

Có bốn loại thang đo được dùng với các biến số: thang đo định danh (nominal scale), thang đo thứ bậc (ordinal scale), thang đo khoảng (interval scale), và thang đo tỉ lệ (ratio scale). Các thang đo này được xác định dựa vào thông tin giá trị được gán cho biến số.

a. Thang đo định danh

Được dùng để chỉ các thuộc tính. Các thuộc tính này được mã hoá bởi các con số dùng để phân loại đối tượng, giữa các con số không có giá trị hơn kém.

Thí dụ: khi khảo sát giới tính, 1 được dùng để chỉ nữ, 2 được dùng để chỉ nam.

b. Thang đo thứ bậc

Được dùng để chỉ các thứ bậc của các đơn vị quan sát. Sự chênh lệch giữa các biểu hiện không nhất thiết phải bằng nhau.

Thí dụ: khảo sát điều kiện làm việc của công nhân, người ta ước lượng mức độ độc hại của chất amiăng (asbestos) đối với công nhân: (1) thấp, (2) trung bình, (3) cao.

c. Thang đo khoảng và thang đo tỉ lệ

Dùng khi các biến được đo ở những khoảng cách đều nhau, chẳng hạn nhiệt độ tính theo độ Celsius (thang đo khoảng) hoặc chiều cao tính theo cm (thang đo tỉ lệ). Giữa hai loại thang đo này có các điểm khác biệt:

- Thang đo tỉ lệ có giá trị zero thật. Thí dụ chiều cao bằng 0 nghĩa là không có chiều cao, trong khi nhiệt độ 0°C không có nghĩa là không có nhiệt độ.

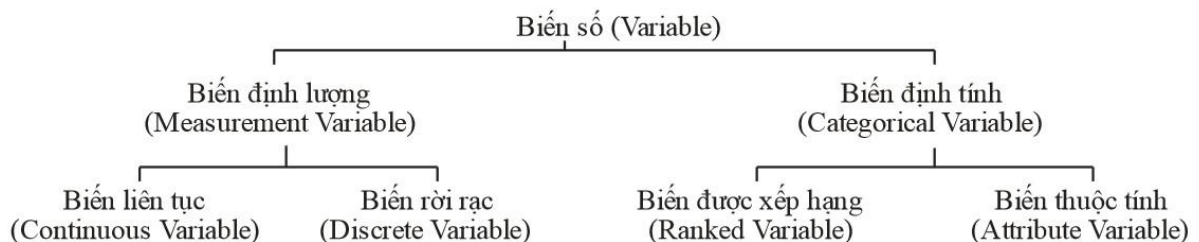
- Khi một biến được đo bằng thang đo tỉ lệ, sự so sánh tỉ lệ giữa hai số là có ý nghĩa. Thí dụ một cây 140cm cao gấp đôi cây 70cm nhưng một lò nung 300°C không nóng gấp đôi lò nung ở 150°C .

Việc sử dụng thang đo thường phụ thuộc vào phương pháp hoặc công cụ đo hơn là thuộc tính. Cùng một thuộc tính có thể được đo bằng các thang khác nhau. Chẳng hạn tuổi có thể được đo theo năm (thang tỉ lệ), hoặc được chia thành 3 nhóm trẻ, trung niên, già (thang thứ bậc)...

4. Các loại biến số trong sinh học (Variables in Biology)

Quan sát và đo lường các hiện tượng là điều cần bản cho tất cả các nghiên cứu khoa học. Các hiện tượng mà ta muốn quan sát được gọi là các biến số (variables), mỗi lĩnh vực nghiên cứu đều có biến số riêng. Biến (Variate) là những đại lượng có thể mang các giá trị khác nhau.

Có thể phân loại các biến số như sau:



a. Biến định lượng (Measurement Variable)

Là những biến mà giá trị của chúng có thể được biểu hiện dưới dạng số và có thể đo đạc. Có hai loại biến định lượng là biến liên tục (Continuous Variable) và biến không liên tục (Discontinuous Variable).

- **Biến liên tục** là biến (về lý thuyết) có một số giá trị xác định nằm giữa hai điểm cố định. Chẳng hạn giữa hai chiều dài 1,5 cm và 1,6 cm có vô số các giá trị có thể đo được. Rất nhiều biến được nghiên cứu trong sinh học là biến liên tục. Thí dụ chiều cao cây (cm), trọng lượng cơ thể (kg) hoặc pH của đất.

- **Biến không liên tục** còn được gọi là biến rời rạc (Discrete Variable = Meristic Variable) là những biến chỉ có những giá trị xác định, không có các giá trị trung gian. Chẳng hạn số đốt trong phần phụ của một loài côn trùng có thể là 4, 5 hoặc 6 nhưng không bao giờ là 4,3 hoặc $5\frac{1}{2}$. Các thí dụ về biến rời rạc là số lượng lá cây, số cá thể con trong một lứa đẻ.

b. Biến định danh/định tính (Categorical Variable)

Là những biến mà giá trị của chúng không thể biểu hiện dưới dạng số thực sự nhưng có thể sắp xếp theo loại.

- **Biến được xếp hạng** (Ranked Variable): Trong một thí nghiệm, người ta có thể ghi nhận *thứ tự nở* của 10 con nhộng mà không lưu ý đến thời điểm nở của mỗi con. Trong trường hợp này các dữ liệu được mã hoá dưới dạng biến được xếp hạng. Thí dụ: chấm điểm mức độ dễ đẻ của bò 1 = không cần can thiệp, 2 = cần can thiệp một ít, 3 = cần bác sĩ thú y; đánh giá mức độ nghiêm trọng của bệnh từ 1 – 5... Trong những trường hợp này, mỗi số được gán cho một loại chứ không phải là thang đo số liệu vì sự khác biệt giữa điểm 1 và 2 không nhất thiết giống với sự khác biệt giữa điểm 2 và 3...

- **Biến thuộc tính** (Attribute Variable = Nominal Variable): Các biến không thể đo đạc, xếp hạng nhưng có thể được biểu hiện về tính chất được gọi là biến thuộc tính. Các thuộc tính có thể là các đặc điểm như trắng hoặc đen, sống hoặc chết, kiểu gen, loại tế bào bạch cầu... Khi các biến thuộc tính được kết hợp với tần số, chúng được gọi là dữ liệu liệt kê (enumeration data) và có thể xử lý thống kê được. Thí dụ: khảo sát 80 con chuột người ta thấy có 4 con lông đen, số còn lại có lông xám. Dữ liệu liệt kê về màu lông chuột có thể sắp xếp như sau:

Màu lông chuột	Tần số (Frequency)
Đen	4
Xám	76
Tổng số chuột	80

5. Độ đúng (Accuracy) và độ chính xác (Precision) của dữ liệu

Độ đúng là trường hợp giá trị tính toán hoặc đo đạc gần với giá trị thực nhất. **Độ chính xác** là trường hợp giá trị đo đạc của các lần lặp lại gần giống nhau nhất. Một cái cân bị lệch nhưng nhạy có thể cho ra số đo không đúng nhưng có độ chính xác. Ngược lại, một cái cân không nhạy có thể cho ra số đo không chính xác. Nếu không có sự sai lệch của dụng cụ đo, độ chính xác sẽ dẫn đến độ đúng, do đó ta cần tập trung hơn vào độ chính xác.

Chương 2. MÔ TẢ VÀ TRÌNH BÀY DỮ LIỆU

Giai đoạn cơ bản và sớm nhất trong khoa học là giai đoạn mô tả. Nếu các sự kiện không được mô tả chính xác thì ta không thể phân tích được chúng. Việc ứng dụng thống kê trong sinh học cũng đi theo xu hướng này. Nếu chúng ta muốn tìm hiểu về mối liên hệ giữa chiều cao của bố mẹ và các người con thì trước tiên ta phải có công cụ thích hợp để đo và xác định chiều cao của từng thành viên trong gia đình này.

Sau khi đã thực hiện các quan sát và thu thập các dữ liệu cho nghiên cứu, công việc đầu tiên là thiết lập bảng dữ liệu dùng để phân tích thống kê. Tiếp đó là trình bày mô tả tóm tắt các dữ liệu đã được thu thập, hoặc bằng các bảng biểu, hoặc bằng các loại đồ thị, sao cho người đọc có thể rút ra được những thông tin cần thiết.

1. MÔ TẢ & TRÌNH BÀY DỮ LIỆU BẰNG BẢNG

Trong phần này các dữ liệu dùng để minh họa được trích từ Digitalis Investigation Group (DIG 1997). Trong bảng 2.1 là các dữ liệu của 40 bệnh nhân được điều trị thử bằng Digoxin để đánh giá hiệu quả và mức độ an toàn của thuốc này trong điều trị bệnh rối loạn nhịp tim (DIG40).

Bảng 2.1. Dữ liệu thử nghiệm lâm sàng Digoxin của 40 bệnh nhân

ID	Treatment	Age	Race	Sex	BMI	SCr	SBP
2289	0	76	1	1	30.586	1.7	130
6745	0	45	1	1	22.85	1.398	130
1322	1	45	1	2	43.269	0.9	115
538	1	31	1	1	27.025	1.159	120
999	1	47	1	2	30.506	1.386	120
3103	0	60	1	1	29.867	1.091	140
1954	1	77	1	1	26.545	1.307	140
5750	1	76	1	1	39.837	1.455	140
1109	0	68	1	2	27.532	1.534	144
4787	1	46	1	1	28.662	1.307	140
666	0	65	1	1	28.058	2	120
6396	0	83	1	1	26.156	1.489	116
5753	1	75	1	1	37.59	1.3	138
1882	0	50	1	1	25.712	1.034	140
5663	0	59	2	1	27.406	1.705	152
6719	1	34	1	1	20.426	1.886	116
4995	0	55	1	1	19.435	1.6	150
4055	0	71	1	1	22.229	1.261	100
4554	1	58	1	2	28.192	1.352	130
2217	1	65	1	1	23.739	1.614	170
896	0	50	1	1	27.406	1.3	140
5368	1	38	1	1	30.853	0.9	134
3403	0	55	1	2	21.79	1.17	130

1426	0	70	1	1	19.04	1.25	150
764	1	63	2	2	28.731	0.9	122
5668	0	74	1	1	29.024	1.227	116
1653	1	63	1	1	28.399	1.1	105
1254	1	73	1	1	26.545	1.3	144
2312	0	78	2	1	22.503	2.682	104
2705	1	66	1	2	28.762	0.9	150
4181	0	44	2	2	26.37	1.148	124
3641	0	64	1	1	21.228	0.9	130
2439	1	49	1	1	15.204	1.307	140
3640	0	79	1	1	18.957	2.239	150
6646	0	61	1	1	27.718	1.659	128
787	0	58	2	2	27.369	0.909	100
5407	1	50	1	2	24.176	1	130
5001	1	70	1	1	19.044	1.2	110
4375	0	61	1	1	32.079	1.273	128
4326	0	65	1	1	29.34	1.2	170

Trong bảng 2.1, các dữ liệu gốc được nhập vào worksheet của chương trình thống kê Minitab (hoặc của Excel). Bảng gồm 8 cột ứng với 8 biến:

ID: mã số của bệnh nhân

Treatment group: nhóm nghiệm thức (lô thí nghiệm) 0 = placebo; 1 = digoxin (placebo là một chất vô hại được dùng như thuốc – trong trường hợp này 0 được xem là lô đối chứng; 1 là lô thí nghiệm).

Age: tuổi được tính bằng năm

Race: sắc tộc 1 = da trắng; 2 = da màu.

Sex: giới tính 1 = nam; 2 = nữ

BMI (Body Mass Index): chỉ số thể trọng = trọng lượng (kg)/chiều cao(m²)

SCr (Serum Creatinine): Creatinine huyết thanh(mg/dL)

SBP (Systolic Blood Pressure): Huyết áp tối đa (mmHg)

1. Bảng tần số một chiều (one-way frequencytable)

Trình bày kết quả quan sát từng mức độ của mỗi biến.

Thí dụ bảng 2.2 trình bày giới tính của 40 bệnh nhân từ dữ liệu gốc đã được nhập trong bảng 2.1. Từ bảng 2.2 ta dễ dàng nhận thấy có ¾ bệnh nhân là nam giới.

Bảng 2.2 Tần số giới tính của 40 bệnh nhân trong DIG40

Giới tính	Số bệnh nhân	Tỉ lệ %
Nam	30	75.0
Nữ	10	25.0
Tổng	40	100.0

Các biến được dùng trong bảng tần số có thể là biến định tính hoặc biến định lượng. Khi trình bày biến liên tục, các giá trị của chúng thường được nhóm lại theo loại.

Thí dụ tuổi thường được xếp loại thành nhóm 10 năm. Bảng 2.3 trình bày tần số của nhóm tuổi ở 40 bệnh nhân trong bảng 2.1.

Bảng 2.3. Tần số độ tuổi của 40 bệnh nhân trong DIG40

Độ tuổi	Số bệnh nhân	Tỉ lệ %
Dưới 40	3	7.5
40 – 49	6	15.0
50 – 59	8	20.0
60 – 69	11	27.5
70 – 79	12	30.0
Tổng	40	100.0

Có hơn phân nửa số bệnh nhân ở độ tuổi từ 60 trở lên. Cần lưu ý là tỉ lệ % có thể được làm tròn nhưng phải bảo đảm tổng là 100%. Đồng thời tiêu đề của bảng cũng phải cung cấp đủ thông tin cho người đọc hiểu được bảng.

2. Bảng tần số hai chiều (Two-way frequency table)

Thường được sử dụng nhiều hơn vì chúng chỉ ra được mối liên hệ giữa các biến. Bảng 2.4 trình bày mối liên hệ giữa giới tính và chỉ số thể trọng (BMI), trong đó BMI đã được chia thành 4 nhóm: ốm ($BMI < 18.5$), bình thường ($18.5 < BMI < 25$), mập ($25 < BMI < 30$), và béo phì ($BMI > 30$).

Bảng 2.4. Chỉ số thể trọng (BMI) và giới tính của 40 bệnh nhân trong DIG40

BMI	Giới tính		Tổng
	Nam	Nữ	
< 18.5 (ốm)	1 (3.3%)	0 (0.0%)	1 (2.5%)
18.5 – 24.9 (bình thường)	10 (33.3%)	2 (20.0%)	12 (30.0%)
25.0 – 29.9 (mập)	14 (46.7%)	6 (60.0%)	20 (50.0%)
≥ 30.0 (béo phì)	5 (16.7%)	2 (20.0%)	7 (17.5%)
Tổng	30	10	40

Có nhiều cách để trình bày dữ liệu dạng bảng biểu. Trong các bảng 2.2, 2.3,

2.4 chúng ta trình bày cả số lượng và tỉ lệ %.. Bảng 2.5 trình bày các đặc điểm của 200 bệnh nhân trong bộ dữ liệu gốc DIG200.

Bảng 2.5 Các đặc điểm cơ bản của 200 bệnh nhân trong DIG200

Các đặc điểm		Tỉ lệ % (n = 200)
Giới tính	Nam	73
	Nữ	27
Sắc tộc	Da trắng	86.5
	Da màu	13.5
Tuổi	Dưới 40	3.5
	40 – 49	11.5
	50 – 59	25
	60 – 69	33
	70 trở lên	26
Chỉ số thể trọng	Ốm (<18.5)	1.5
	Bình thường (18.5 – 24.9)	37.5
	Mập (25 – 29.9)	42.5
	Béo phì (≥ 30)	18.5

Ngoài tần số, các dữ liệu khác cũng có thể trình bày dưới dạng bảng. Thí dụ bảng 2.6 trình bày chi phí y tế tính bằng % GDP của 3 nước Hoa Kỳ, Canada và Anh trong thời gian từ 1960 đến 2000.

Bảng 2.6 Chi phí y tế (%GDP) trong giai đoạn từ 1960 đến 2000

Năm	Canada	UK	USA
1960	5.4	3.9	5.1
1965	5.6	4.1	6.0
1970	7.0	4.5	7.0
1975	7.0	5.5	8.4
1980	7.1	5.6	8.8
1985	8.0	6.0	10.6
1990	9.0	6.0	12.0
1995	9.2	7.0	13.4
2000	9.2	7.3	13.3

II. MÔ TẢ & TRÌNH BÀY DỮ LIỆU BẰNG BIỂU ĐỒ

Chương trình Minitab có thể được dùng để mô tả dữ liệu bằng nhiều dạng biểu đồ khác nhau, bao gồm:

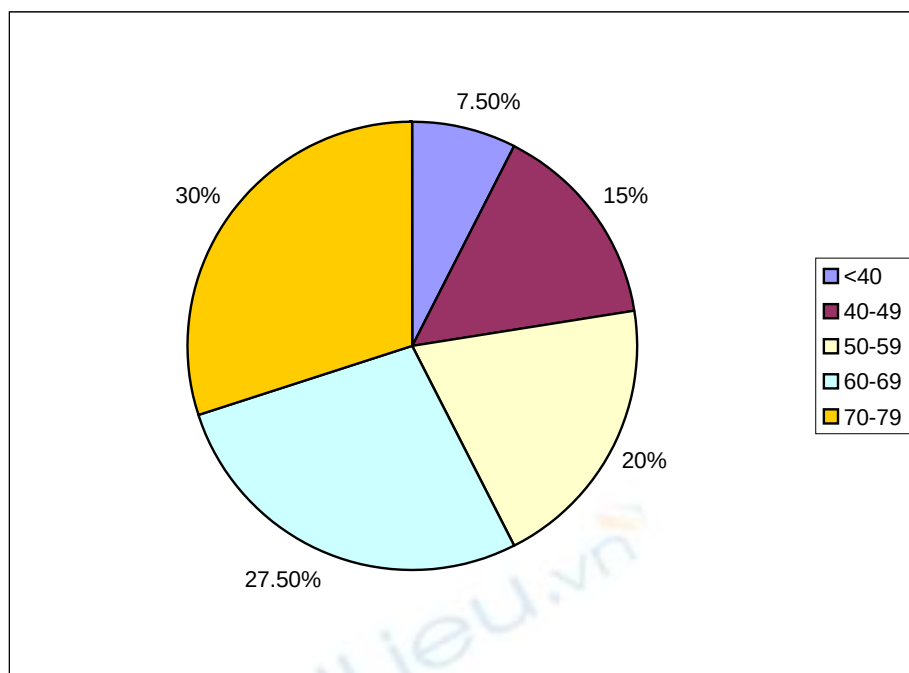
- Một chiều: histograms, boxplots, dotplots...
- Hai chiều: scatter plots, matrixplots...
- Ba chiều: contour plots, surfaceplots.

Trong giáo trình này chúng ta làm quen với một số dạng biểu đồ thường được dùng để mô tả các dữ liệu và cách sử dụng chúng cho phù hợp tùy theo mục đích.

1. Pie chart (biểu đồ hình quạt = biểu đồ hình tròn):

Thường được dùng để so sánh các giá trị dữ liệu dưới dạng tỉ lệ %.

Dữ liệu của bảng 2.3 được trình bày dưới dạng biểu đồ Pie (Hình 2.1)



Hình 2.1 Biểu đồ Pie dạng 2D

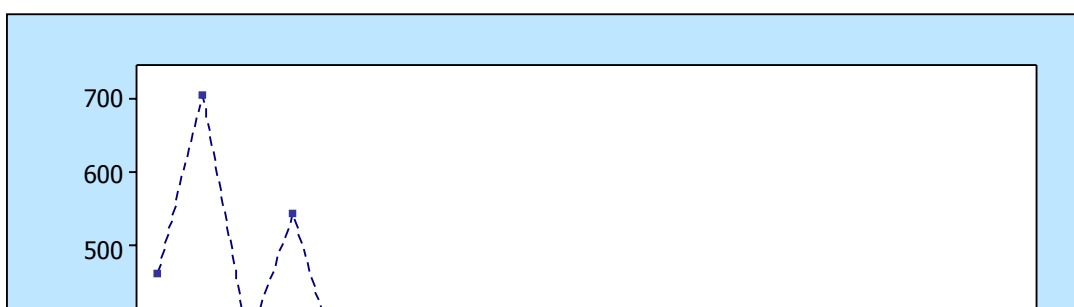
2. Time Series Plot

Thường dùng để *biểu diễn sự thay đổi của các dữ liệu theo thời gian*.

Thí dụ: Theo dõi nồng độ của Hg trong 20 năm ở hai vị trí khác nhau của Địa Trung hải. Ở mỗi vị trí, 45 mẫu của loài *P. oceanica* được thu thập ở độ sâu 10m và mang về phòng thí nghiệm để xác định nồng độ Hg. Nồng độ Hg trung bình (ng/ g trọng lượng khô) của các mẫu ở mỗi vị trí được ghi nhận theo từng năm như trong bảng sau:

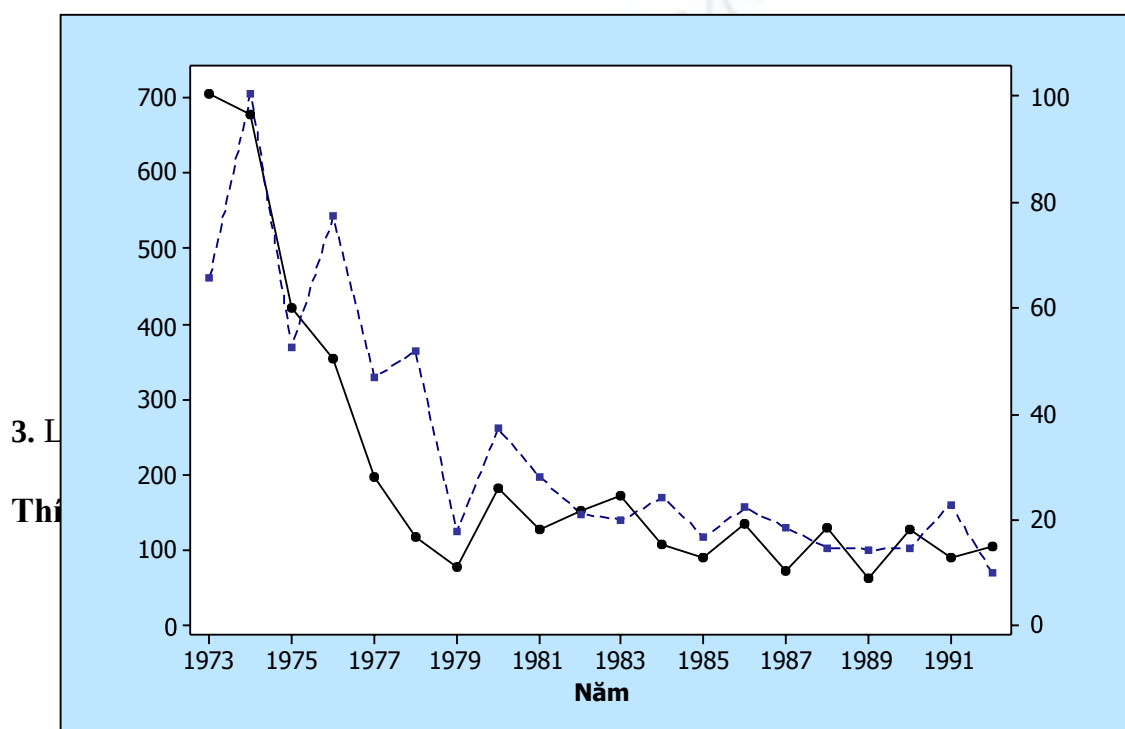
Nồng độ thủy ngân					
Year	Site 1	Site 2	Year	Site 1	Site 2
1992	14.80	70.20	1982	21.50	147.80
1991	12.90	160.50	1981	18.20	197.70
1990	18.00	102.80	1980	25.80	262.10
1989	8.70	100.30	1979	11.00	123.30
1988	18.30	103.10	1978	16.50	363.90
1987	10.30	129.00	1977	28.10	329.40
1986	19.30	156.20	1976	50.50	542.60
1985	12.70	117.60	1975	60.10	369.90
1984	15.20	170.60	1974	96.70	705.10
1983	24.60	139.60	1973	100.40	462.00

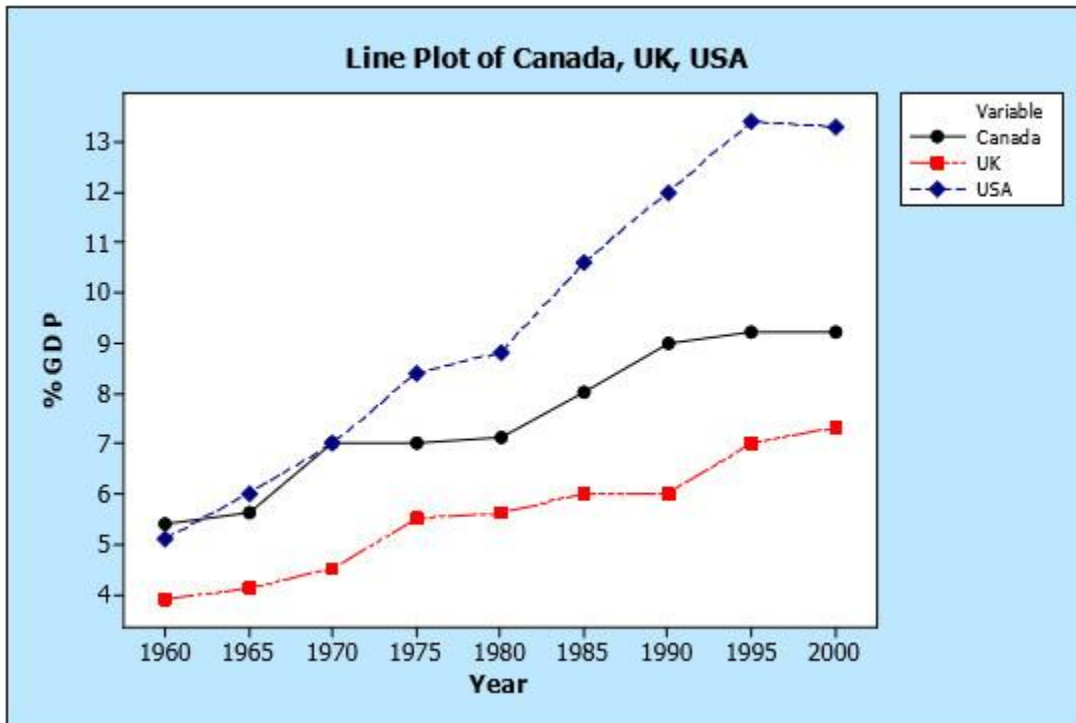
Dùng dữ liệu trong bảng trên để xây dựng biểu đồ line :



3. Line Graph

Thường dùng để so sánh dữ liệu của hai hoặc nhiều nhóm. Thí dụ: Dùng dữ liệu trong bảng 2.6 để xây dựng biểu đồ line :

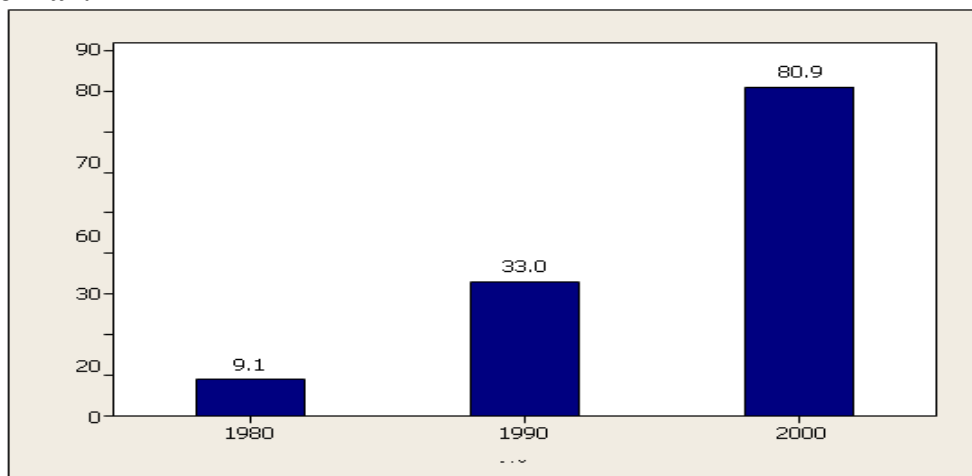




4. BarChart

Để so sánh các giá trị của các chuỗi dữ liệu người ta có thể dùng các biểu đồ Bar Chart hoặc Area Graph.

Thí dụ 1. Ở Mỹ, số người tham gia tổ chức bảo vệ sức khỏe (HMOs = Health Maintenance Organization) là 9.1 triệu năm 1980, 33.0 triệu năm 1990 và 80.9 triệu năm 2000. Thông tin này được trình bày trong hình 2.5 bằng biểu đồ Bar.



Hình: 2.5 Biểu đồ Bar về số người tham gia HMOs theo năm

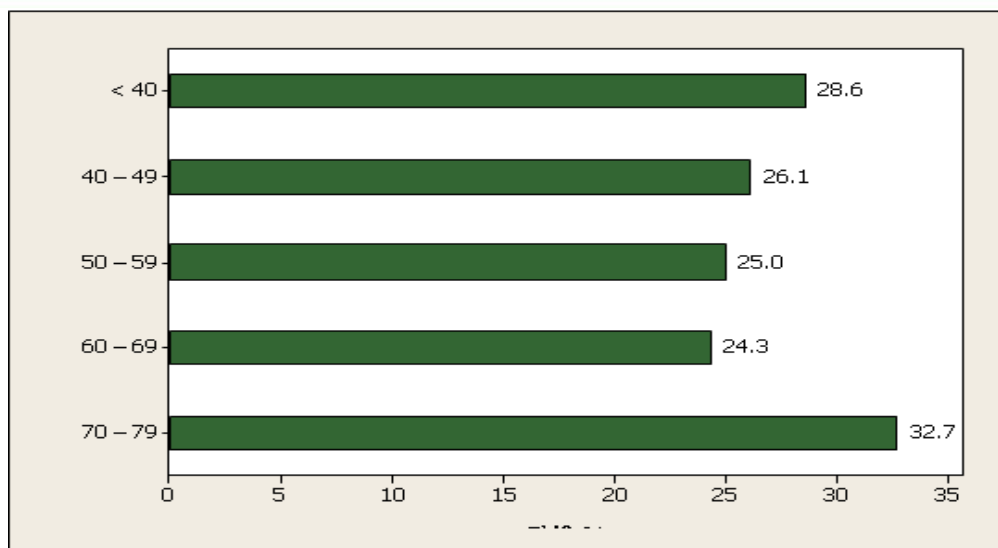
Biểu đồ Bar có thể dùng trình bày các dữ liệu phức tạp hơn, chẳng hạn các dữ liệu trong các bảng 2 chiều hoặc 3 chiều.

Thí dụ 2. Bảng 2.7 trình bày tỉ lệ người mắc bệnh tiểu đường theo 5 độ tuổi.

Dữ liệu này được trình bày bằng biểu đồ Bar như trong hình 2.6.

Bảng 2.7 Tỷ lệ người mắc bệnh tiểu đường theo 5 độ tuổi

Độ tuổi	Tỷ lệ %
< 40	28.6
40 – 49	26.1
50 – 59	25.0
60 – 69	24.3
70 – 79	32.7



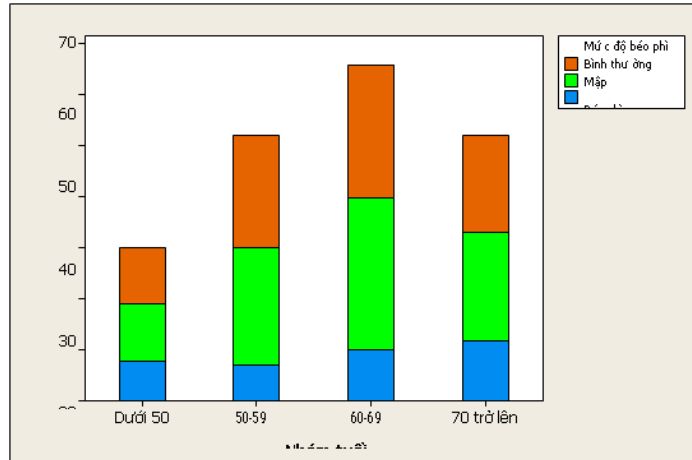
Hình: 2.6. Biểu đồ Bar trình bày tỷ lệ người bệnh tiểu đường ở mỗi nhóm tuổi

Khi các biến trong bảng 2 chiều có nhiều hơn hai mức độ cho mỗi biến, chúng ta có thể dùng biểu đồ segmentbar.

Thí dụ 3 các dữ liệu (số người và tỷ lệ %) về mối liên quan giữa sự béo phì và tuổi được trình bày trong một bảng 3 x 4 (bảng 2.8)

Bảng 2.8 Mối liên quan giữa mức độ béo phì và nhóm tuổi

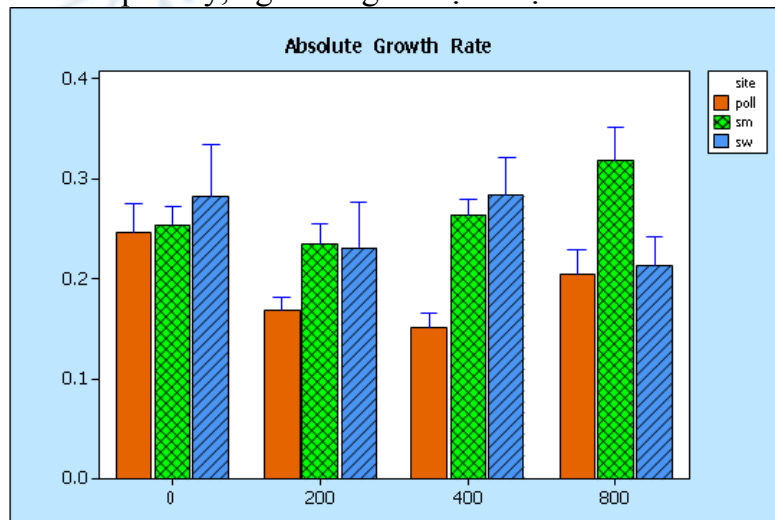
Độ béo phì	Nhóm tuổi (trong ngoặc đơn là tỷ lệ %)			
	Dưới 50	50 – 59	60 – 69	70 trở lên
Bình thường	11 (36.6)	22 (42.3)	26 (39.4)	19 (36.5)
Mập	11 (36.6)	23 (44.2)	30 (45.5)	21 (40.4)
Béo phì	8 (26.7)	7 (13.5)	10 (15.2)	12 (23.1)
Tổng	30	52	66	52



5. Interval Plot

Dùng để biểu diễn giá trị trung bình cùng với khoảng tin cậy hoặc sai số. Biểu đồ này vừa minh họa cho các số đo trung tâm, vừa cho thấy biến động của dữ liệu.

Thí dụ: Nhằm khảo sát sự tăng trưởng của trùng Enchytraeid, người ta thu mẫu ở 3 địa điểm khác nhau và đem nuôi trong môi trường có nồng độ là 0, 200, 400, và 800 mg Zn/kg. Số đốt thân được đếm 2 ngày/lần trong một tháng. Kết quả được lưu trong worksheet Enchytraeid. MTW. Từ kết quả này, người ta ghi nhận được biểu đồ như trong hình 2.9



6. Histogram

Biểu đồ này tương tự biểu đồ Bar nhưng chỉ dùng cho các biến được phân lớp. Biểu đồ cung cấp cho chúng ta hình ảnh về sự phân bố của dữ liệu. Các giá trị của biến được nhóm thành từng lớp (còn gọi là bin) thường có độ rộng bằng nhau. Số biến trong mỗi lớp được chỉ bởi độ cao của cột. Trước khi thiết lập biểu đồ, các dữ liệu phải được sắp xếp lại trong bảng tần số 1 chiều. Các dữ liệu trong bảng 2.9 được dùng để thiết lập biểu đồ Histogram.

Bảng 2.9 Tần số huyết áp tối đa (mmHg) của 199 bệnh nhân

Giá trị	Tần số	Giá trị	Tần số	Giá trị	Tần số	Giá trị	Tần số	Giá trị	Tần số	Giá trị	Tần số
85	1	105	1	116	8	128	3	138	1	150	12
90	5	106	2	118	5	130	23	139	2	152	3
95	2	108	2	120	25	131	1	140	26	155	1
96	1	110	16	122	4	132	2	142	1	160	3
100	14	112	1	124	4	134	1	144	3	162	1
102	1	114	5	125	3	135	2	145	1	165	1
104	2	115	2	126	1	136	1	148	1	170	5

Trước khi vẽ biểu đồ, ta phải trả lời được các câu hỏi sau:

- (1) Cần phải chia dữ liệu thành bao nhiêu lớp?
- (2) Độ rộng (khoảng) của mỗi lớp là bao nhiêu?

Như vậy ta phải tính toán số lớp và độ rộng của mỗi lớp để xây dựng bảng phân bố tần số (bảng 2.10).

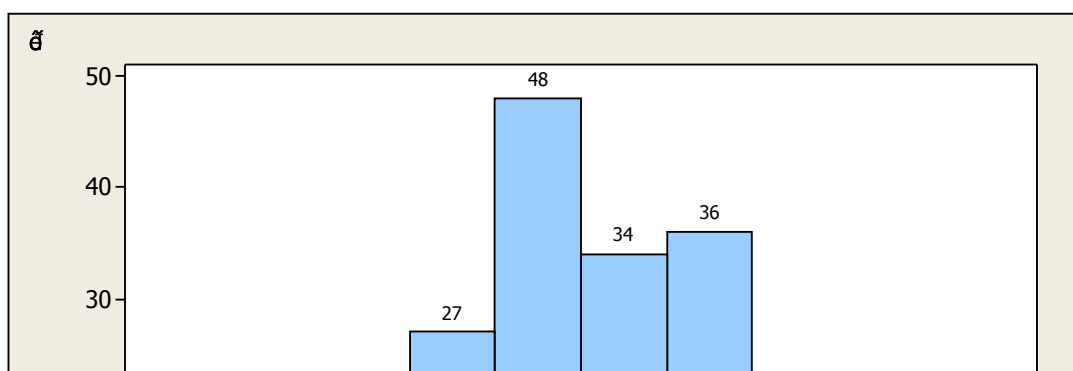
(1) Số lượng khoảng có liên quan đến số lượng biến. Nói chung có thể chia thành từ 5 – 15 khoảng. Cỡ mẫu càng nhỏ thì số lượng khoảng càng ít. Một phương pháp được đề nghị bởi Sturges và Scott (1979) là dùng $\log_2 n + 1$ để tính số khoảng (trong đó n là số lượng biến quan sát được). Trong bảng 2.9 có 199 biến, ta cần tìm giá trị của $\log_2 199 + 1$. Giá trị này là 8.64, có thể làm tròn là 9, nghĩa là có thể dùng 9 khoảng để xây dựng biểu đồ.

(2) Tính độ rộng của mỗi lớp (class interval) = $(X_{\max} - X_{\min}) / (\log_2 n + 1)$ Vì $(170 - 85) / 8.64 = 85 / 8.64 = 9.84$, làm tròn thành 10mmHg.

Bảng 2.10 Phân bố tần số huyết áp tối đa của các dữ liệu trong bảng 2.9

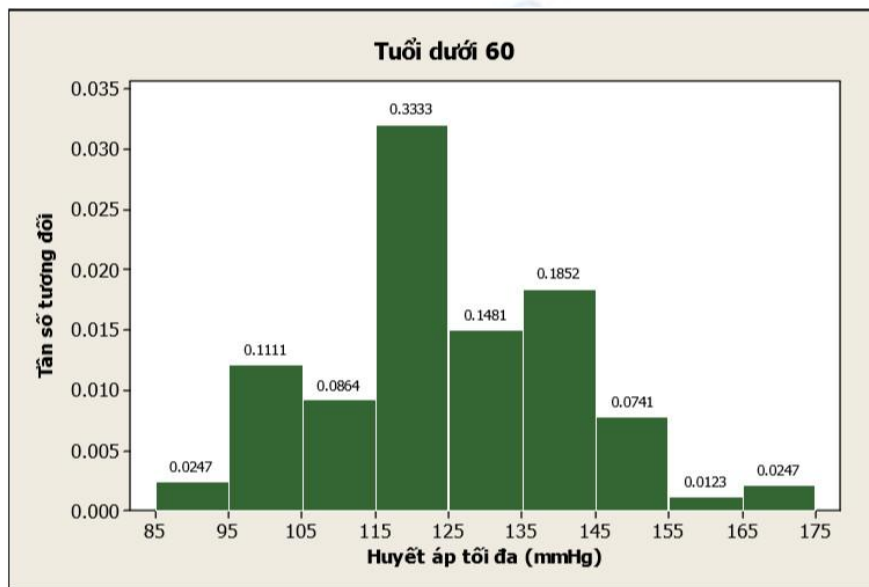
Lớp (Bin)	Khoảng lớp (Class Width)	Tần số (Frequency)	Tần suất (Relative Frequency)	Tần suất tích lũy (Cumulative Relative Frequency)	Tần số tích lũy (Cumulative Frequency)
1	85 – 95	6	3.02	3.02	6
2	95 -105	20	10.05	13.07	26
3	105 – 115	27	13.57	26.63	53
4	115 – 125	48	24.12	50.75	101
5	125 – 135	34	17.09	67.84	135
6	135 – 145	36	18.09	85.93	171
7	145 – 155	17	8.54	94.47	188
8	155 – 165	5	2.51	96.98	193
9	165 – 175	6	3.02	100.00	199
Tổng		199	100.00		

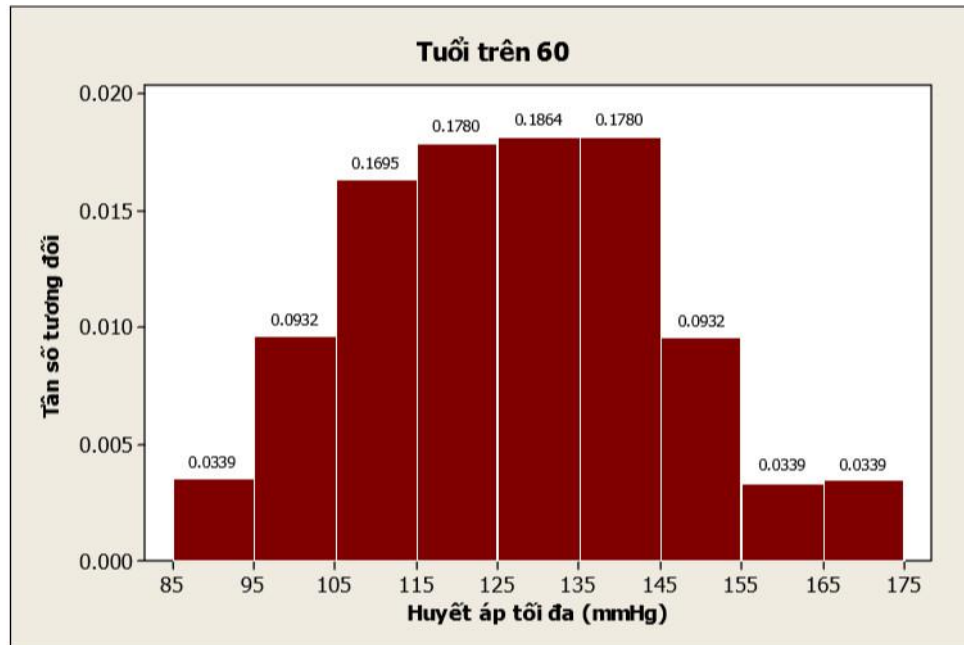
Cần lưu ý là 85 – 95 bao gồm tất cả các giá trị từ 85 – 95 nhưng không bao gồm 95 nghĩa là $85 \leq y < 95$.



Hình: 2.10 Biểu đồ histogram về huyết áp tối đa của 199 bệnh nhân

Có thể dùng biểu đồ histogram để so sánh hai nhóm dữ liệu. Thí dụ dùng bộ dữ liệu DIG200 ta xây dựng hai biểu đồ histogram để so sánh sự phân bố huyết áp tối đa của hai nhóm tuổi dưới 60 và trên 60. Hình 2.11 trình bày hai biểu đồ của hai nhóm này. Cần lưu ý là ta sử dụng tần số tương đối (tần suất) thay vì tần số do dữ liệu của hai nhóm này khác nhau (cỡ mẫu khác nhau): chỉ có 81 bệnh nhân dưới 60 tuổi trong khi có tới 118 bệnh nhân trên 60 tuổi.





Hình 2.11 Biểu đồ histogram về phân bố huyết áp tối đa theo nhóm tuổi

7. Stem-and-Leaf Plots

Biểu đồ này tương tự histogram, chỉ khác ở chỗ chúng trình bày giá trị dữ liệu thay vì dùng các thanh (cột). Biểu đồ stem-and-leaf thường chỉ dùng cho các nhóm dữ liệu nhỏ. Thí dụ xây dựng biểu đồ huyết áp tối đa đo được của 40 bệnh nhân trong bộ dữ liệu DIG40.

8. Dotplot (biểu đồ điểm)

Một dotplot trình bày sự phân bố của một biến liên tục. Hình 2.12 dưới đây là dotplot giúp chúng ta so sánh sự phân bố của biến liên tục là huyết áp tối đa qua một biến định tính là tuổi được chia thành hai nhóm: dưới 60 và trên 60 tuổi (DIG 40). Tương tự stem-and-leaf plot, dotplot được dùng cho các nhóm dữ liệu nhỏ.

9 Scatterplots

Scatterplot là biểu đồ hai chiều (two- dimension) được sử dụng rất phổ biến để trình bày mối quan hệ giữa hai nhóm dữ liệu thu được trong thí nghiệm. Đặc biệt chúng được dùng rất hiệu quả với các biến liên tục.

Hình 2.13 là biểu đồ scatter dùng khảo sát tương quan giữa mức creatinine trong huyết thanh với huyết áp tối đa của 40 bệnh nhân (DIG40). Trong biểu đồ mỗi vòng tròn tượng trưng cho một giá trị creatinine huyết thanh và huyết áp tối đa của một bệnh nhân. Chẳng hạn, vòng tròn trên cùng phía tay trái của biểu đồ chỉ bệnh nhân thứ hai (ID = 2312) với creatinine huyết thanh là 2.682 mg/dL và huyết áp tối đa là 104mmHg.

Có nhiều khả năng là giá trị của cả hai biến đều giống nhau ở một số bệnh nhân. Khảo sát cẩn thận các dữ liệu trong bảng 2.1 (DIG40) cho thấy có 3 bệnh nhân (ID = 4787, 1954, 2439) đều có creatinine là 1.307 mg/dL và huyết áp tối đa là 140 mmHg. Chúng được biểu diễn bằng các vòng tròn chồng lên nhau thành một.