Chương này tập trung vào cơ chế và cấu trúc các tóm tắt thống kê và đồ thị. Chúng tôi sử dụng phần mềm thống kê để tạo ra các tóm tắt và đồ thị được trình bày trong chương và cuốn sách này. Tuy nhiên, vì đây có thể là lần đầu tiên bạn tiếp xúc với những khái niệm này, chúng tôi dành thời gian trong chương này để trình bày chi tiết cách tạo chúng. Nắm vững nội dung được trình bày trong chương này sẽ rất quan trọng để hiểu các phương pháp và kỹ thuật được giới thiệu trong phần còn lại của cuốn sách.

**Trang 40**

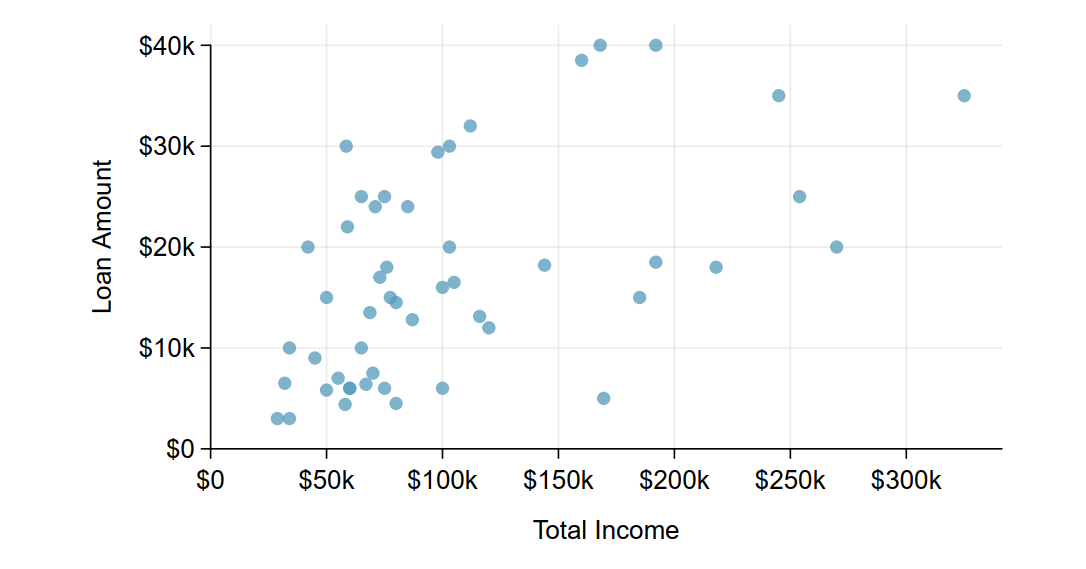
**2.1 Kiểm tra dữ liệu kiểu số**

Trong phần này chúng ta sẽ tìm hiểu các kỹ thuật tóm tắt các biến số. Ví dụ: hãy xem xét biến **loan\_amount** từ tập dữ liệu **loan50**, đại diện cho quy mô khoản vay cho tất cả 50 khoản vay trong tập dữ liệu. Biến này là số vì chúng ta có thể thảo luận một cách hợp lý về sự khác biệt số của quy mô của hai khoản vay. Mặt khác, mã vùng và mã zip không phải là số, mà chúng là các biến phân loại.

Trong suốt phần này và phần tiếp theo, chúng tôi sẽ áp dụng các phương pháp này bằng cách sử dụng bộ dữ liệu loan50 và quận, đã được giới thiệu trong Phần 1.2. Nếu bạn muốn xem lại các biến từ một trong hai tập dữ liệu, hãy xem **Hình 1.3 và 1.5.**

**2.1.1 Dùng biểu đồ phân so sánh dữ liệu**

Biểu đồ phân tán cung cấp chế độ xem dữ liệu theo từng trường hợp cho hai biến số. Trong **Hình 1.8** trên **trang 16**, biểu đồ phân tán được sử dụng để kiểm tra tỷ lệ sở hữu nhà so với phần nhỏ của các đơn vị nhà ở là một phần của tài sản nhiều đơn vị (ví dụ: căn hộ) trong tập dữ liệu của quận. Một biểu đồ phân tán khác được thể hiện trong **Hình 2.1**, so sánh tổng thu nhập của một người đi vay (tổng thu nhập) và số tiền họ đã vay (số tiền vay) cho tập dữ liệu **loan50**. Trong bất kỳ biểu đồ phân tán nào, mỗi điểm đại diện cho một trường hợp duy nhất. Vì có 50 trường hợp trong **loan50**, có 50 điểm trong **Hình 2.1**.

  
**Hình 2.1**: Biểu đồ phân tán của tổng thu nhập so với số tiền cho vay cho tập dữ liệu **loan50**.

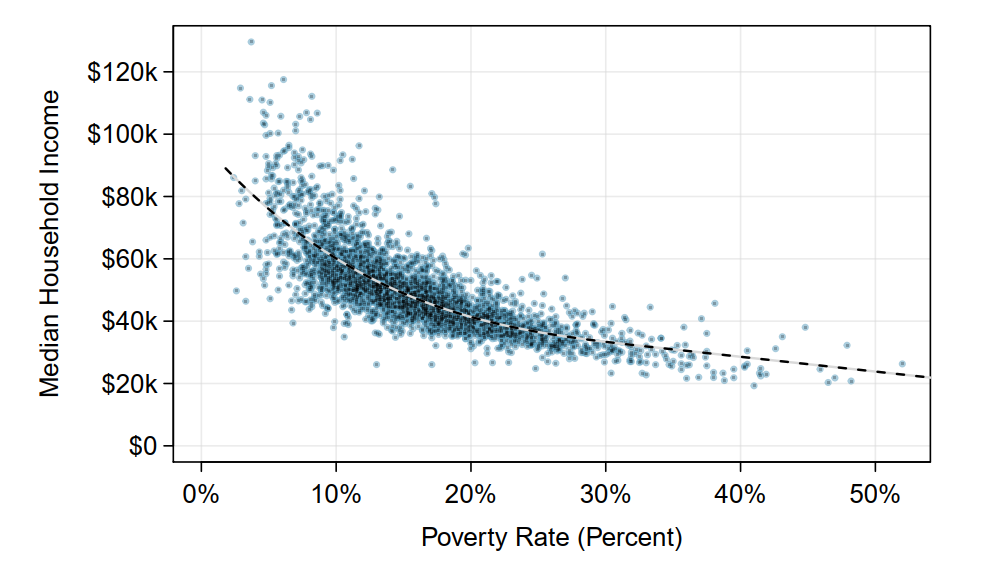
Nhìn vào **Hình 2.1,** chúng ta thấy rằng có rất nhiều người đi vay với thu nhập dưới 100.000 USD ở phía bên trái của biểu đồ, trong khi có một số ít những người đi vay có thu nhập trên 250.000 USD.

**VÍ DỤ 2.1**

Hình 2.2 cho thấy một biểu đồ thu nhập trung bình của hộ gia đình so với tỷ lệ nghèo ở 3.142 quận. Có thể nói gì về mối quan hệ giữa các biến này?

**Trang 41**

Mối quan hệ rõ ràng là phi tuyến tính, như được đánh dấu bằng đường đứt nét. Điều này khác với các biểu đồ phân tán trước đây mà chúng ta đã thấy, biểu đồ này cho thấy các mối quan hệ không cho thấy nhiều, nếu có, độ cong trong xu hướng.

 **Hình 2.2**: Biểu đồ phân tán của thu nhập trung bình của hộ gia đình so với tỷ lệ nghèo cho tập dữ liệu của quận. Một mô hình thống kê cũng phù hợp với dữ liệu và được hiển thị dưới dạng đường đứt nét.

**HƯỚNG DẪN THỰC HÀNH 2.2**

Biểu đồ phân tán tiết lộ điều gì về dữ liệu và chúng hữu ích như thế nào?

1 Câu trả lời có thể khác nhau. Scatterplots rất hữu ích trong việc nhanh chóng phát hiện ra các mối liên quan có liên quan đến các biến số, cho dù các mối quan hệ đó ở dạng xu hướng đơn giản hay liệu các mối quan hệ đó phức tạp hơn.

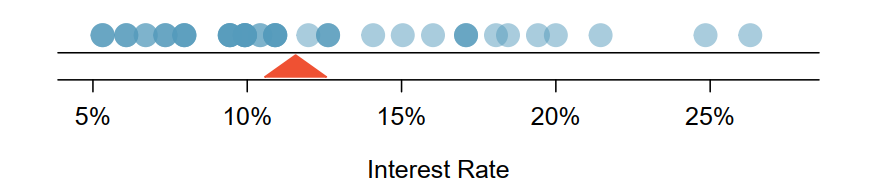
**HƯỚNG DẪN THỰC HÀNH 2.3**Mô tả hai biến có thể có liên kết hình móng ngựa trong biểu đồ phân tán (∩ hoặc \_).

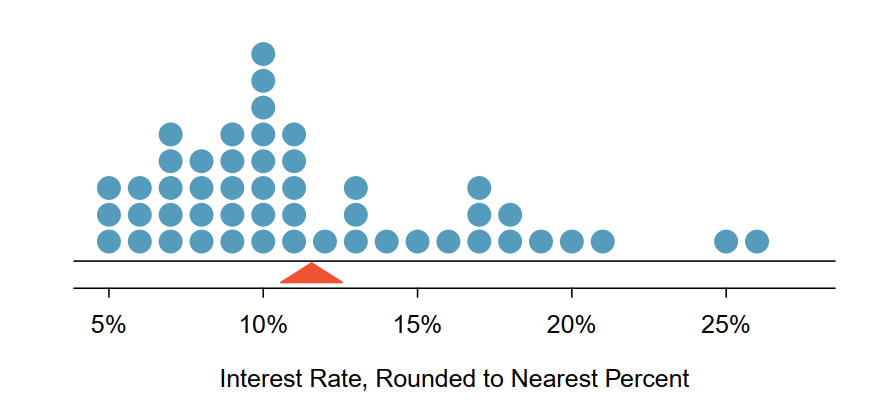
2 Hãy xem xét trường hợp trục tung của bạn đại diện cho điều gì đó “tốt” và trục hoành đại diện cho điều gì đó chỉ tốt ở mức độ vừa phải. Sức khỏe và tiêu thụ nước phù hợp với mô tả này: chúng ta cần một số nước để tồn tại, nhưng tiêu thụ quá nhiều và nó trở nên độc hại và có thể giết chết một người.

**Trang 42**

**2.1.2 Đồ thị điểm và giá trị trung bình**

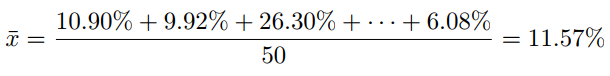
Đôi khi hai biến là một quá nhiều: chỉ một biến có thể được quan tâm. Trong những trường hợp này, một biểu đồ điểm cung cấp những điều cơ bản nhất về hiển thị. Biểu đồ điểm là biểu đồ phân tán một biến; một ví dụ sử dụng lãi suất của **loan50** được thể hiện trong **Hình 2.3**. Một phiên bản xếp chồng của biểu đồ chấm này được thể hiện trong **Hình 2.4.**

  
**Hình 2.3**: Biểu đồ lãi suất cho **loan50**. Giá trị trung bình của phân phối được hiển thị dưới dạng hình tam giác màu đỏ.



**Hình 2.4:** Biểu đồ lãi suất dạng chấm xếp chồng lên nhau cho tập dữ liệu **loan50**. Tỷ lệ đã được làm tròn đến phần trăm gần nhất trong biểu đồ này và giá trị trung bình của phân phối được hiển thị dưới dạng hình tam giác màu đỏ

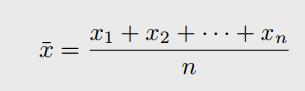
**Giá trị trung bình**, thường được gọi là **giá trị trung bình cộng**, là một cách phổ biến để đo lường trung tâm của một phân phối dữ liệu. Để tính lãi suất trung bình, chúng tôi cộng tất cả các lãi suất và chia cho số lần quan sát:



Giá trị trung bình của mẫu thường được gắn nhãn **¯x**. Chữ x đang được sử dụng như một trình giữ chỗ chung cho biến số lãi suất, lãi suất và thanh trên chữ x biểu thị chúng ta đang xem lãi suất trung bình, đối với 50 khoản vay này là 11,57%. Sẽ rất hữu ích nếu coi giá trị trung bình là điểm cân bằng của phân phối và nó được thể hiện dưới dạng hình tam giác trong **Hình 2.3 và 2.4.**

**MEAN**

Giá trị trung bình của mẫu có thể được tính bằng tổng các giá trị quan sát được chia cho số lần quan sát:



Trong đó x1, x2 ,. . . , xn đại diện cho n giá trị quan sát được.

**HƯỚNG DẪN THỰC HÀNH 2.4**

Kiểm tra phương trình cho giá trị trung bình. X1 tương ứng với cái gì? Và x2? Bạn có thể suy ra một ý nghĩa chung cho những gì xi có thể đại diện?

*x1 tương ứng với lãi suất của khoản vay đầu tiên trong mẫu (10,90%), x2 với lãi suất của khoản vay thứ hai (9,92%) và xi tương ứng với lãi suất của khoản vay thứ i trong tập dữ liệu. Ví dụ: nếu i = 4, thì chúng tôi đang kiểm tra x4, đề cập đến quan sát thứ tư trong tập dữ liệu.*

**HƯỚNG DẪN THỰC HÀNH 2.5**

N là bao nhiêu trong mẫu khoản vay này?

*Cỡ mẫu là n = 50.*

Tập dữ liệu loan50 đại diện cho một mẫu từ một nhóm lớn hơn các khoản vay được thực hiện thông qua Câu lạc bộ cho vay. Chúng tôi có thể tính giá trị trung bình cho dân số này theo cách tương tự như giá trị trung bình của mẫu. Tuy nhiên, trung bình dân số có một nhãn đặc biệt: µ. Ký hiệu µ là chữ cái Hy Lạp mu và đại diện cho giá trị trung bình của tất cả các quan sát trong quần thể. Đôi khi, một chỉ số con, chẳng hạn như x, được sử dụng để đại diện cho biến nào mà trung bình tổng thể đề cập đến, ví dụ: µx. Thông thường, việc đo lường chính xác trung bình dân số là quá đắt, vì vậy chúng tôi thường ước tính µ bằng cách sử dụng trung bình mẫu, ¯x.

**Trang 43**

**HƯỚNG DẪN THỰC HÀNH 2.6**

Lãi suất trung bình của tất cả các khoản vay trong dân số có thể được ước tính bằng cách sử dụng dữ liệu mẫu. Dựa trên mẫu của 50 khoản vay, ước tính hợp lý của µx, lãi suất trung bình cho tất cả các khoản vay trong tập dữ liệu đầy đủ là bao nhiêu?

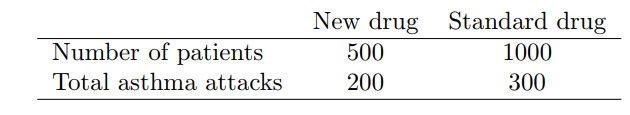
Giá trị trung bình của mẫu, 11,57%, cung cấp một ước tính sơ bộ về µx. Mặc dù nó không hoàn hảo nhưng đây là dự đoán tốt nhất của chúng tôi về lãi suất trung bình của tất cả các khoản vay trong dân số đang được nghiên cứu.

Trong Chương 5 và hơn thế nữa, chúng tôi sẽ phát triển các công cụ để mô tả độ chính xác của các ước tính điểm như giá trị trung bình của mẫu. Như bạn có thể đoán, các ước tính điểm dựa trên các mẫu lớn hơn có xu hướng chính xác hơn so với các ước tính dựa trên các mẫu nhỏ hơn.

*6 Độ lệch có thể nhìn thấy trong cả ba biểu đồ, mặc dù biểu đồ chấm phẳng là ít hữu ích nhất. Biểu đồ và biểu đồ chấm xếp chồng lên nhau là những hình ảnh hữu ích để xác định độ lệch.*

**HƯỚNG DẪN THỰC HÀNH 2.7**

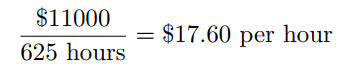
Giá trị trung bình rất hữu ích vì nó cho phép chúng tôi bán lại hoặc chuẩn hóa một số liệu thành một thứ gì đó dễ hiểu và dễ so sánh hơn. Đưa ra 2 ví dụ trong đó giá trị trung bình hữu ích để so sánh.

1. Chúng tôi muốn hiểu liệu một loại thuốc mới có hiệu quả hơn trong việc điều trị các cơn hen suyễn so với loại thuốc tiêu chuẩn hay không. Một thử nghiệm với 1500 người lớn được thiết lập, trong đó 500 người nhận được loại thuốc mới và 1000 người nhận được loại thuốc tiêu chuẩn trong nhóm đối chứng:  


So sánh số lượng thô của 200 đến 300 cơn hen suyễn sẽ cho thấy rằng loại thuốc mới tốt hơn, nhưng đây là một giả thuyết của quy mô nhóm không cân đối. Thay vào đó, chúng ta nên xem xét số cơn hen suyễn trung bình của mỗi bệnh nhân trong mỗi nhóm:

New drug: 200*/*500 = 0*.*4 Standard drug: 300*/*1000 = 0*.*3

2. Emilio đã mở một xe bán thức ăn vào năm ngoái, nơi anh ấy bán bánh mì kẹp thịt, và công việc kinh doanh của anh ấy đã ổn định trong 3 tháng qua. Trong khoảng thời gian 3 tháng đó, anh ấy đã kiếm được 11.000 đô la trong khi làm việc 625 giờ. Thu nhập trung bình hàng giờ của Emilio cung cấp một thống kê hữu ích để đánh giá xem liệu doanh nghiệp của anh ấy, ít nhất là từ góc độ tài chính, có xứng đáng hay không:



Bằng cách biết mức lương trung bình hàng giờ của mình, Emilio giờ đây đã đưa thu nhập của mình vào một đơn vị tiêu chuẩn dễ so sánh với nhiều công việc khác mà anh ấy có thể cân nhắc.

*7Lãi suất cho các khoản vay cá nhân.*

**HƯỚNG DẪN THỰC HÀNH 2.8**

Giả sử chúng ta muốn tính thu nhập trung bình trên mỗi người ở Mỹ. Để làm như vậy, trước tiên chúng ta có thể nghĩ đến việc lấy giá trị trung bình của thu nhập bình quân đầu người trên 3.142 quận trong tập dữ liệu quận. Cách tiếp cận tốt hơn sẽ là gì?*8 Có thể có hai nhóm chiều cao hiển thị trong tập dữ liệu: một của học sinh và một của người lớn. Đó là, dữ liệu có thể là hai phương thức.*

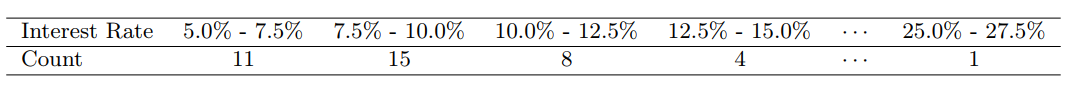
Bộ dữ liệu quận đặc biệt ở chỗ mỗi quận thực sự đại diện cho nhiều người dân. Nếu chúng ta chỉ đơn giản là tính trung bình trên biến thu nhập, chúng ta sẽ đối xử bình đẳng với các quận có 5.000 và 5.000.000 cư dân trong các phép tính. Thay vào đó, chúng ta nên tính tổng thu nhập cho mỗi quận, cộng tổng tất cả các quận, sau đó chia cho số người trong tất cả các quận. Nếu chúng tôi hoàn thành các bước này với dữ liệu của quận, chúng tôi sẽ thấy rằng thu nhập bình quân đầu người ở Hoa Kỳ là 30.861 đô la. Nếu chúng tôi tính toán mức trung bình đơn giản của thu nhập bình quân đầu người giữa các quận, kết quả sẽ chỉ là 26.093 đô la!

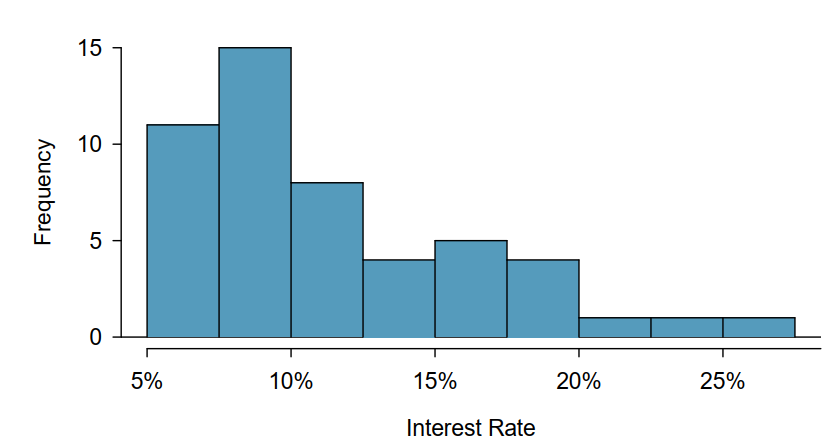
Ví dụ này sử dụng cái được gọi là giá trị trung bình có trọng số. Để biết thêm thông tin về chủ đề này, hãy xem phần bổ sung trực tuyến sau về phương tiện có trọng số http://www.openintro.org/redirect.php?go=stat\_wtd\_mean

**Trang 44**

**2.1.3 Biểu đồ Histogram và hình dạng**

Biểu đồ điểm hiển thị giá trị chính xác cho mỗi lần quan sát. Điều này hữu ích cho các tập dữ liệu nhỏ, nhưng chúng có thể trở nên khó đọc với các mẫu lớn hơn. Thay vì hiển thị giá trị của mỗi quan sát, chúng tôi thích nghĩ giá trị thuộc về một thùng. Ví dụ: trong tập dữ liệu loan50, chúng tôi đã tạo một bảng đếm số lượng các khoản vay có lãi suất từ 5,0% đến 7,5%, sau đó là số lượng các khoản vay có lãi suất từ 7,5% đến 10,0%, v.v. Các quan sát rơi vào ranh giới của một thùng (ví dụ: 10,00%) được phân bổ cho thùng phía dưới. Bảng này được thể hiện trong **Hình 2.5**. Các số đếm được xếp chồng này được vẽ dưới dạng các thanh trong **Hình 2.6** thành biểu đồ được gọi là biểu đồ, giống như một phiên bản được xếp chồng nhiều hơn của biểu đồ chấm xếp chồng được thể hiện trong **Hình 2.4.**

 **Hình 2.5: Số lượng cho dữ liệu lãi suất được phân bổ.**



**Hình 2.6: Biểu đồ lãi suất. Sự phân bố này bị lệch mạnh về bên phải.**

Biểu đồ cung cấp một cái nhìn về mật độ dữ liệu. Các thanh cao hơn thể hiện nơi dữ liệu tương đối phổ biến hơn. Ví dụ, có nhiều khoản vay có lãi suất từ 5% đến 10% hơn các khoản vay có tỷ lệ từ 20% đến 25% trong tập dữ liệu. Các thanh này giúp bạn dễ dàng xem mật độ của dữ liệu thay đổi như thế nào so với lãi suất.

Biểu đồ đặc biệt thuận tiện để hiểu hình dạng của phân phối dữ liệu. Hình 2.6 cho thấy hầu hết các khoản vay có lãi suất dưới 15%, trong khi chỉ một số ít các khoản vay có lãi suất trên 20%. Khi dữ liệu đi lệch sang phải theo cách này và có đuôi bên phải dài hơn, hình dạng được cho là lệch phải.5

Các tập dữ liệu có đặc điểm ngược lại - đuôi dài, mỏng hơn về bên trái - được cho là lệch trái. Chúng tôi cũng nói rằng sự phân bố như vậy có một cái đuôi dài bên trái. Các tập dữ liệu hiển thị dấu vết gần bằng nhau theo cả hai hướng được gọi là đối xứng.

**XÁC ĐỊNH ĐỘ XIÊN BẰNG CHIỀU DÀI ĐUÔI**Khi dữ liệu đi theo một hướng, phân phối có một đuôi dài. Nếu một bản phân phối có đuôi dài bên trái, nó bị lệch trái. Nếu một phân phối có đuôi dài bên phải, nó bị lệch phải

**Trang 45**

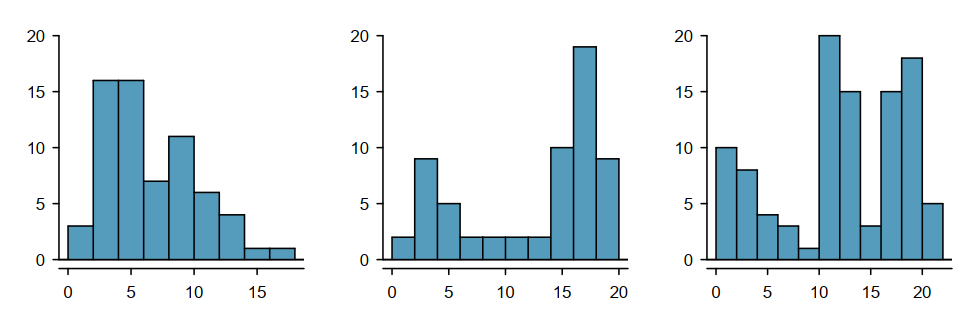
**HƯỚNG DẪN THỰC HÀNH 2.9**Hãy xem các biểu đồ chấm trong Hình 2.3 và 2.4. Bạn có thể thấy độ lệch trong dữ liệu không? Có dễ dàng nhìn thấy độ lệch trong biểu đồ này hay các biểu đồ dấu chấm không?

**HƯỚNG DẪN THỰC HÀNH 2.10**Ngoài giá trị trung bình (vì nó được gắn nhãn), bạn có thể thấy gì trong các ô chấm mà bạn không thể thấy trong biểu đồ?

Ngoài việc xem phân bố có lệch hay đối xứng hay không, biểu đồ có thể được sử dụng để xác định các chế độ. Một chế độ được thể hiện bằng một đỉnh nổi bật trong phân phối. Chỉ có một đỉnh nổi bật trong biểu đồ số tiền cho vay.

Định nghĩa về mode đôi khi được dạy trong các lớp toán là giá trị có nhiều lần xuất hiện nhất trong tập dữ liệu. Tuy nhiên, đối với nhiều tập dữ liệu trong thế giới thực, thông thường không có quan sát nào có cùng giá trị trong tập dữ liệu, làm cho định nghĩa này không thực tế trong phân tích dữ liệu.

Hình 2.7 cho thấy các biểu đồ có một, hai hoặc ba đỉnh nổi bật. Các phân phối như vậy được gọi là đơn phương thức, hai phương thức và đa phương thức, tương ứng. Bất kỳ phân phối nào có nhiều hơn 2 đỉnh nổi bật được gọi là đa phương thức. Lưu ý rằng có một đỉnh nổi bật trong phân phối đơn phương với đỉnh thứ hai kém nổi bật hơn không được tính vì nó chỉ khác với các thùng lân cận của nó qua một vài lần quan sát.

  
**Hình 2.7**: Chỉ đếm các đỉnh nổi bật, các phân bố là (từ trái sang phải) đơn phương thức, hai phương thức và đa phương thức. Lưu ý rằng chúng tôi đã nói rằng âm mưu bên trái là cố ý không theo phương thức. Điều này là do chúng tôi đang đếm các đỉnh nổi bật, không chỉ bất kỳ đỉnh nào.

**VÍ DỤ 2.11  
Hình 2.6** chỉ cho thấy một chế độ nổi bật trong lãi suất. Phân phối là đơn phương thức, hai phương thức hay đa phương thức?

Không phương thức. Hãy nhớ rằng uni là viết tắt của 1 (nghĩ là xe hai bánh). Tương tự, bi là viết tắt của 2 (nghĩ rằng xe đạp). Chúng tôi hy vọng một chu trình đa vòng sẽ được phát minh để hoàn thành sự tương tự này.

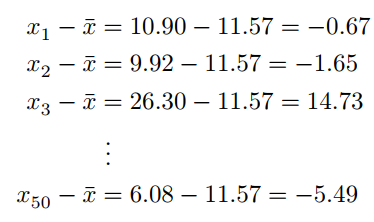
**HƯỚNG DẪN THỰC HÀNH 2.12**Các phép đo chiều cao của học sinh nhỏ tuổi và giáo viên người lớn tại một trường tiểu học K-3 đã được thực hiện. Bạn mong đợi bao nhiêu chế độ trong tập dữ liệu chiều cao này?

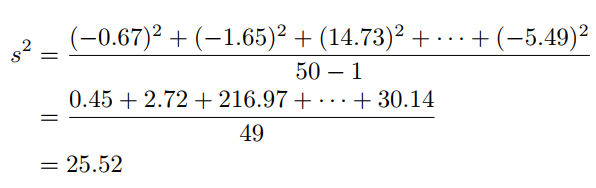
Tìm kiếm chế độ không phải là tìm câu trả lời rõ ràng và chính xác về số lượng chế độ trong một bản phân phối, đó là lý do tại sao nổi bật không được xác định chặt chẽ trong cuốn sách này. Phần quan trọng nhất của cuộc kiểm tra này là để hiểu rõ hơn về dữ liệu của bạn.

**2.1.4 Phương sai và độ lệch chuẩn**

**Trang 46**

Giá trị trung bình được giới thiệu như một phương pháp để mô tả trung tâm của một tập dữ liệu và sự thay đổi trong dữ liệu cũng rất quan trọng. Ở đây, chúng tôi giới thiệu hai thước đo độ biến thiên: phương sai và độ lệch chuẩn. Cả hai điều này đều rất hữu ích trong phân tích dữ liệu, mặc dù các công thức của chúng hơi tẻ nhạt khi tính toán bằng tay. Độ lệch chuẩn dễ hiểu hơn cả hai và nó mô tả một cách đại khái khoảng cách giữa quan sát điển hình so với giá trị trung bình.

Chúng tôi gọi khoảng cách của một quan sát so với độ lệch trung bình của nó. Dưới đây là độ lệch cho các lần quan sát thứ nhất, thứ 2, thứ 3 và thứ 50 trong biến lãi suất:  


Nếu chúng ta bình phương các độ lệch này và sau đó lấy giá trị trung bình, kết quả bằng phương sai mẫu, được ký hiệu là *s*2:  


Chúng tôi chia cho n - 1, thay vì chia cho n, khi tính toán phương sai của mẫu; có một số sắc thái toán học ở đây, nhưng kết quả cuối cùng là việc làm này làm cho thống kê này đáng tin cậy và hữu ích hơn một chút.

Lưu ý rằng bình phương độ lệch làm được hai điều. Đầu tiên, nó làm cho các giá trị lớn tương đối lớn hơn nhiều, được thấy bằng cách so sánh (−0,67) 2, (−1,65) 2, (14,73) 2 và (−5,49) 2. Thứ hai, nó loại bỏ bất kỳ dấu hiệu tiêu cực nào.

Độ lệch chuẩn được định nghĩa là căn bậc hai của phương sai:

Mặc dù thường bị bỏ qua, một chỉ số phụ của x có thể được thêm vào phương sai và độ lệch chuẩn, tức là s2x và sx, nếu nó hữu ích như một lời nhắc nhở rằng đây là phương sai và độ lệch chuẩn của các quan sát được đại diện bởi *x*1, *x*2, ..., *xn*.

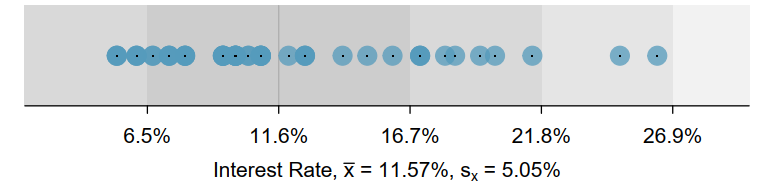
**PHƯƠNG SAI VÀ ĐỘ LỆCH CHUẨN**

Phương sai là khoảng cách bình phương trung bình từ giá trị trung bình. Độ lệch chuẩn là căn bậc hai của phương sai. Độ lệch chuẩn rất hữu ích khi xem xét dữ liệu được phân phối bao xa so với giá trị trung bình.

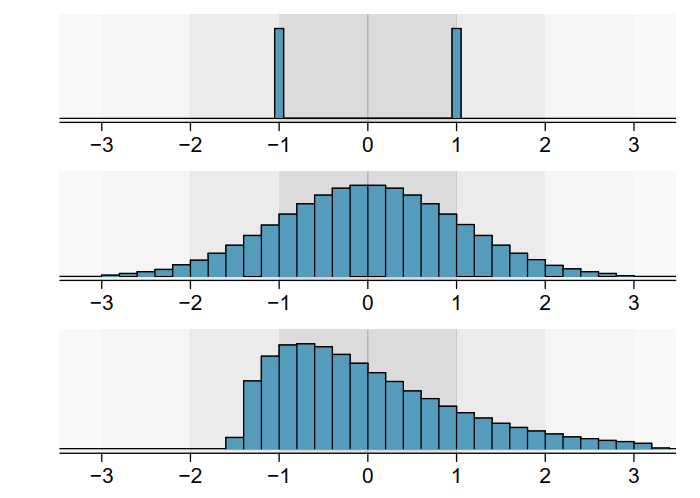
Độ lệch chuẩn thể hiện độ lệch điển hình của các quan sát so với giá trị trung bình. Thông thường, khoảng 70% dữ liệu sẽ nằm trong một độ lệch chuẩn của giá trị trung bình và khoảng 95% sẽ nằm trong hai độ lệch chuẩn. Tuy nhiên, như trong Hình 2.8 và 2.9, những tỷ lệ phần trăm này không phải là những quy tắc nghiêm ngặt.

Giống như giá trị trung bình, các giá trị tổng thể cho phương sai và độ lệch chuẩn có các ký hiệu đặc biệt: σ2 cho phương sai và σ cho độ lệch chuẩn. Ký hiệu σ là chữ cái Hy Lạp sigma.

**Trang 47**



**Hình 2.8**: Đối với biến lãi suất, 34 trong số 50 khoản vay (68%) có lãi suất nằm trong 1 độ lệch chuẩn của giá trị trung bình và 48 trong số 50 khoản vay (96%) có lãi suất nằm trong 2 độ lệch chuẩn. Thông thường, khoảng 70% dữ liệu nằm trong 1 độ lệch chuẩn của giá trị trung bình và 95% nằm trong 2 độ lệch chuẩn, mặc dù điều này không phải là một quy tắc cứng.

  
**Hình 2.9**: Ba phân bố dân số rất khác nhau có cùng giá trị trung bình µ = 0 và độ lệch chuẩn σ = 1.

**HƯỚNG DẪN THỰC HÀNH 2.13**

Ở **trang 45**, khái niệm về hình dạng của một phân phối đã được giới thiệu. Một mô tả tốt về hình dạng của một phân phối phải bao gồm phương thức và liệu phân phối có đối xứng hay lệch sang một bên hay không. Sử dụng **Hình 2.9** làm ví dụ, hãy giải thích tại sao mô tả như vậy lại quan trọng.

**Hình 2.9** cho thấy ba phân phối trông khá khác nhau, nhưng tất cả đều có cùng giá trị trung bình, phương sai và độ lệch chuẩn. Sử dụng phương thức, chúng ta có thể phân biệt giữa âm mưu đầu tiên (hai phương thức) và hai âm mưu cuối cùng (đơn phương thức). Sử dụng độ lệch, chúng ta có thể phân biệt giữa âm mưu cuối cùng (lệch phải) và hai ô đầu tiên. Trong khi một bức tranh, giống như biểu đồ, kể một câu chuyện hoàn chỉnh hơn, chúng ta có thể sử dụng phương thức và hình dạng (đối xứng / lệch) để mô tả thông tin cơ bản về phân bố

**VÍ DỤ 2.14**

Mô tả sự phân bố của biến lãi suất bằng cách sử dụng biểu đồ trong Hình 2.6. Mô tả phải kết hợp trung tâm, sự thay đổi và hình dạng của phân phối và nó phải được đặt trong ngữ cảnh. Cũng cần lưu ý bất kỳ trường hợp đặc biệt bất thường nào.

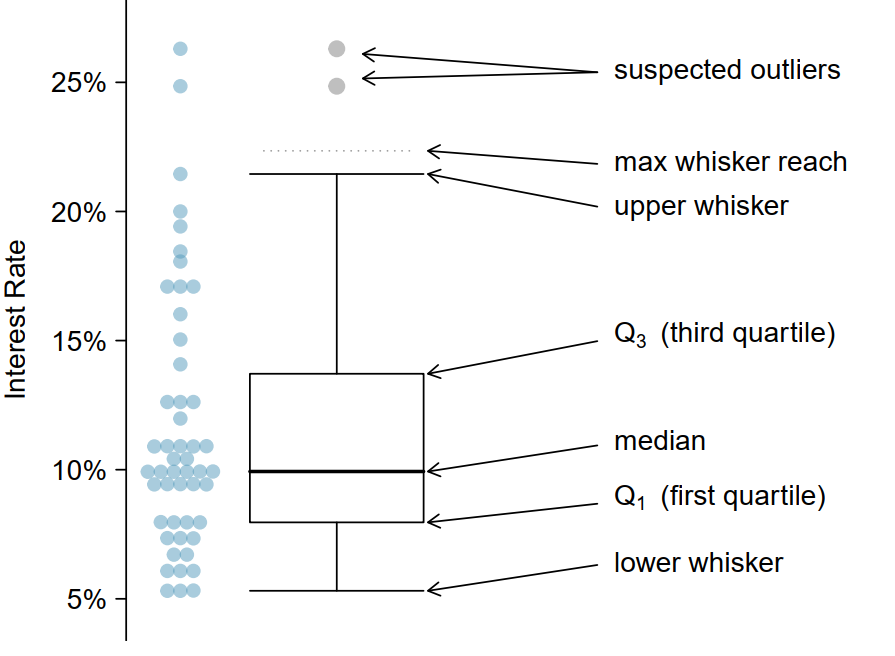
Việc phân bổ lãi suất không theo phương thức và nghiêng về mức cao. Nhiều tỷ lệ giảm xuống gần mức trung bình ở mức 11,57%, và hầu hết nằm trong một độ lệch chuẩn (5,05%) của mức trung bình. Có một vài mức lãi suất đặc biệt lớn trong mẫu là trên 20%.

Trong thực tế, phương sai và độ lệch chuẩn đôi khi được sử dụng như một phương tiện để kết thúc, ở đó “phần cuối” có thể ước tính chính xác độ không đảm bảo liên quan đến thống kê mẫu. Ví dụ, trong **Chương 5,** độ lệch chuẩn được sử dụng trong các phép tính giúp chúng ta hiểu giá trị trung bình của một mẫu thay đổi bao nhiêu từ mẫu này sang mẫu tiếp theo.

**Trang 48**

**2.1.5 Box plots, tứ phân vị và trung vị**

Một biểu đồ hộp tóm tắt một tập dữ liệu bằng cách sử dụng năm số liệu thống kê đồng thời vẽ biểu đồ các quan sát bất thường. **Hình 2.10** cung cấp một biểu đồ chấm dọc cùng với một biểu đồ hộp của biến lãi suất từ tập dữ liệu **loan50**.

  
**Hình 2.10**: Một biểu đồ chấm dọc, trong đó các điểm được xếp chồng lên nhau theo chiều ngang, bên cạnh một ô hộp được dán nhãn cho lãi suất của 50 khoản vay.

Bước đầu tiên trong việc xây dựng biểu đồ hình hộp là vẽ một đường tối biểu thị đường trung bình, chia đôi dữ liệu. Hình 2.10 cho thấy 50% dữ liệu nằm dưới mức trung bình và 50% khác nằm trên mức trung bình. Có 50 khoản vay trong tập dữ liệu (một số chẵn) vì vậy dữ liệu được tách hoàn hảo thành hai nhóm 25. Chúng tôi lấy giá trị trung bình trong trường hợp này là giá trị trung bình của hai quan sát gần nhất với phân vị thứ 50, xảy ra cùng giá trị trong tập dữ liệu này: (9,93% + 9,93%) / 2 = 9,93%. Khi có một số lượng quan sát lẻ, sẽ có chính xác một quan sát chia dữ liệu thành hai nửa và trong trường hợp đó quan sát đó là giá trị trung bình (không cần trung bình).

**TRUNG VỊ: SỐ CÓ VỊ TRÍ TRUNG TÂM**

Nếu dữ liệu được sắp xếp theo thứ tự từ nhỏ nhất đến lớn nhất, trung vị là quan sát ngay chính giữa. Nếu có một số lượng quan sát chẵn, sẽ có hai giá trị ở giữa và giá trị trung vị được lấy làm giá trị trung bình của chúng.

Bước thứ hai trong việc xây dựng một biểu đồ hình hộp là vẽ một hình chữ nhật để thể hiện 50% dữ liệu ở giữa. Tổng chiều dài của hộp, được thể hiện theo chiều dọc trong **Hình 2.10**, được gọi là phạm vi liên phần (viết tắt là IQR). Nó, giống như độ lệch chuẩn, là một thước đo sự thay đổi của dữ liệu. Dữ liệu càng thay đổi, độ lệch chuẩn và IQR càng lớn. Hai ranh giới của hộp được gọi là phần tư thứ nhất (phần trăm thứ 25, tức là 25% dữ liệu nằm dưới giá trị này) và phần tư thứ ba (phần trăm thứ 75), và chúng thường được gắn nhãn tương ứng là Q1 và Q3.

**Trang 49**