

Thống kê trong kinh doanh (Phần 1)

Tác giả chính

ALEXANDER HOLMES, THE UNIVERSITY OF OKLAHOMA

BARBARA ILLOWSKY, DE ANZA COLLEGE

SUSAN DEAN, DE ANZA COLLEGE

Biên dịch

TS. Nguyễn Tuấn Anh (Chủ biên)

ThS. Nguyễn Thị Phương Anh

ThS. Võ Thị Ngọc Hà

TS. Nguyễn Mạnh Hoàng

ThS. Vũ Nhật Phương

ThS. Trần Hoàng Cẩm Tú

ThS. Tạ Hữu Hiến

Introductory Business Statistics 2e

SENIOR CONTRIBUTING AUTHORS

ALEXANDER HOLMES, THE UNIVERSITY OF OKLAHOMA

BARBARA ILLOWSKY, DE ANZA COLLEGE

SUSAN DEAN, DE ANZA COLLEGE

Copyright Disclamier (for Introductory Business Statistics 2e)

OpenStax

Rice University

6100 Main Street MS-375

Houston, Texas 77005

To learn more about OpenStax, visit <https://openstax.org>.
Individual print copies and bulk orders can be purchased through our website.

©2023 Rice University. Textbook content produced by OpenStax is licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). Under this license, any user of this textbook or the textbook contents herein must provide proper attribution as follows:

- If you redistribute this textbook in a digital format (including but not limited to PDF and HTML), then you must retain on every page the following attribution:
“Access for free at openstax.org.”
- If you redistribute this textbook in a print format, then you must include on every physical page the following attribution:
“Access for free at openstax.org.”
- If you redistribute part of this textbook, then you must retain in every digital format page view (including but not limited to PDF and HTML) and on every physical printed page the following attribution:
“Access for free at openstax.org.”
- If you use this textbook as a bibliographic reference, please include
<https://openstax.org/details/books/introductory-business-statistics-2e> in your citation.

For questions regarding this licensing, please contact support@openstax.org.

Trademarks

The OpenStax name, OpenStax logo, OpenStax book covers, OpenStax CNX name, OpenStax CNX logo, OpenStax Tutor name, OpenStax Tutor logo, Connexions name, Connexions logo, Rice University name, and Rice University logo are not subject to the license and may not be reproduced without the prior and express written consent of Rice University.

HARDCOVER BOOK ISBN-13	978-1-711472-
B&W PAPERBACK BOOK ISBN-13	978-1-711472-
DIGITAL VERSION ISBN-13	978-1-961584-
ORIGINAL PUBLICATION YEAR	2023
1 2 3 4 5 6 7 8 9 10 CJP 23	

OPENSTAX

OpenStax provides free, peer-reviewed, openly licensed textbooks for introductory college and Advanced Placement® courses and low-cost, personalized courseware that helps students learn. A nonprofit ed tech initiative based at Rice University, we're committed to helping students access the tools they need to complete their courses and meet their educational goals.

RICE UNIVERSITY

OpenStax is an initiative of Rice University. As a leading research university with a distinctive commitment to undergraduate education, Rice University aspires to path-breaking research, unsurpassed teaching, and contributions to the betterment of our world. It seeks to fulfill this mission by cultivating a diverse community of learning and discovery that produces leaders across the spectrum of human endeavor.

MỤC LỤC

MỤC LỤC	iv
Lời nói đầu (cho bản dịch Tiếng Việt)	1
Preface for Introductory Business Statistics 2e	3
1 LẤY MẪU VÀ DỮ LIỆU	9
Giới thiệu	9
1.1 Định nghĩa thống kê, xác suất và các thuật ngữ chính	9
1.2 Dữ liệu, lấy mẫu và biến thiên trong dữ liệu và lấy mẫu	16
1.3 Thang đo	33
1.4 Thiết kế thực nghiệm và đạo đức nghiên cứu	42
Thuật ngữ chính	45
Tóm tắt chương	47
Tài liệu tham khảo	49
2 THỐNG KÊ MÔ TẢ	50
Giới thiệu	50
2.1 Trình bày dữ liệu	51
2.2 Các số đo vị trí của dữ liệu	75
2.3 Các số đo tâm của dữ liệu	87
2.4 Ký hiệu sigma và tính trung bình số học	93
2.5 Trung bình nhân	94
2.6 Độ lệch và bình quân, trung vị, và yếu vị	96
2.7 Số đo độ phân tán của dữ liệu	98
Thuật ngữ chính	109
Tóm tắt chương	110
Tóm tắt công thức	112
Tài liệu tham khảo	114
3 XÁC SUẤT	116
Giới thiệu	116
3.1 Thuật ngữ	116
3.2 Các sự kiện độc lập và xung khắc	124
3.3 Hai quy tắc cơ bản của xác suất	133
3.4 Bảng tiếp liên	138
3.5 Sơ đồ Venn	150

Thuật ngữ chính	160
Tóm tắt chương	162
Tóm tắt công thức	163
Tài liệu tham khảo	163
4 BIẾN NGẪU NHIÊN RỜI RẠC.....	166
Giới thiệu	166
4.1. Phân phối siêu bội.....	169
4.2. Phân phối nhị thức.....	173
4.3. Phân phối hình học	178
4.4. Phân phối Poisson.....	186
Thuật ngữ chính	195
Tóm tắt chương	197
Tóm tắt công thức	199
Tài liệu tham khảo	201
5 BIẾN NGẪU NHIÊN LIÊN TỤC.....	204
Giới thiệu	204
5.1 Các tính chất của hàm mật độ xác suất	204
5.2 Phân phối đồng nhất.....	209
5.3 Phân phối mũ.....	211
Thuật ngữ chính	228
Tóm tắt chương	229
Tóm tắt công thức	230
Tài liệu tham khảo	231
6 PHÂN PHỐI CHUẨN.....	232
Giới thiệu	232
6.1. Phân phối chuẩn đơn vị	233
6.2 Sử dụng phân phối chuẩn	236
6.3 Ước lượng phân phối nhị thức bằng phân phối chuẩn.....	244
Thuật ngữ chính	249
Tóm tắt chương	249
Tóm tắt công thức	250
Tài liệu tham khảo	251

Lời nói đầu (cho bản dịch Tiếng Việt)

Đây là bản dịch Tiếng Việt (phần 1) của giáo trình **Introductory Business Statistics 2e** (an OpenStax resource, access for free at openstax.org). Bản dịch này phục vụ yêu cầu giảng dạy và học tập môn học thống kê trong các ngành kinh tế - quản trị, chương trình chuẩn quốc tế, chương trình song ngữ, chương trình chất lượng cao tại Trường Đại học Nguyễn Tất Thành.

Giáo trình **Introductory Business Statistics 2e** được lựa chọn sử dụng trong giảng dạy, học tập môn học thống kê, bên cạnh đó bởi ưu điểm của giáo trình là một tài liệu mở, miễn phí, còn do một số lý do như sau:

- Phần lý thuyết thống kê của Giáo trình được trình bày đơn giản, dễ hiểu, không đi sâu vào việc chứng minh công thức toán, mà trọng tâm vào giải thích khái niệm;
- Giáo trình có nhiều ví dụ minh họa, bài tập vận dụng gắn liền với cuộc sống, hoạt động kinh doanh, giúp người đọc không chỉ hiểu rõ được lý thuyết, mà còn thấy được ứng dụng của môn học;
- Giáo trình có nhiều bài tập thực hành, giúp người học vận dụng ngay kiến thức vào giải quyết vấn đề.

Giáo trình **Introductory Business Statistics 2e** được cung cấp trong khuôn khổ OpenStax. OpenStax là một sáng kiến của Đại học Rice (Mỹ), nhằm cung cấp cho người học các giáo trình, học liệu mở. Theo thông tin từ trang web openstax.org (tháng 11/2025), sáng kiến này đã cung cấp hơn 60 đầu sách giáo trình đại học, được sử dụng bởi 7 triệu sinh viên từ 150 quốc gia; 70% trường đại học, cao đẳng Mỹ có sử dụng giáo trình của OpenStax.

Mục đích chính khi biên dịch Giáo trình này là cung cấp cho người học một phiên bản Tiếng Việt để có thể sử dụng song song với bản gốc Tiếng Anh. Một vấn đề quan trọng khi biên dịch là lựa chọn thuật ngữ Tiếng Việt tương đương với các thuật ngữ thống kê Tiếng Anh. Nhóm biên dịch đã thống nhất sử dụng Từ điển Toán học Anh – Việt, xuất bản năm 1976 bởi NXB Khoa học và Kỹ thuật như là tài liệu tham chiếu thuật ngữ chính, đồng thời có tham khảo thêm Từ điển Tiếng Việt của Viện Ngôn ngữ (thường được gọi là từ điển Hoàng Phê, xuất bản bởi NXB Đà Nẵng năm 2003) và Từ điển khoa học – kỹ thuật Anh – Việt của NXB Thế giới (năm 2003). Để giúp người học dễ đối chiếu thuật ngữ thống kê Anh, Việt, thuật ngữ Tiếng Anh được ghi kèm với thuật ngữ Tiếng Việt khi thuật ngữ này xuất hiện lần đầu tiên trong Giáo trình. Tuy vậy, do một số thuật ngữ hiện nay chưa thống nhất, hoặc có nhiều từ khác nhau cho cùng một từ Tiếng Anh nên nhóm biên dịch sử dụng các ghi chú chân trang để cung cấp cho người học thêm thông tin về các thuật ngữ khác (nếu có).

Do mục đích chính là cung cấp phiên bản Tiếng Việt dùng kèm bản gốc Tiếng Anh, nên phần bài tập thực hành không được dịch. Người học sẽ được yêu cầu sử dụng bản gốc để thực hành vận dụng kiến thức.

Bản dịch này (so với bản gốc Tiếng Anh) cũng bổ sung phần sử dụng các hàm của phần mềm Excel và R để tính toán các công thức thống kê. Phần bổ sung này không nhằm hướng dẫn người học toàn diện cách sử dụng phần mềm Excel, R để xử lý thống kê dữ liệu, mà chỉ có mục đích cung cấp thông tin về các hàm mà người học có thể sử dụng để giải các

bài tập vận dụng, thực hành trong Giáo trình, thay vì chỉ đơn thuần sử dụng bàn toán. Nhóm biên dịch cũng giới hạn chỉ giới thiệu các hàm built-in của R và hàm trong package stats.

Trong quá trình biên dịch, Nhóm biên dịch cũng đã điều chỉnh một số thuật ngữ, công thức trong bản gốc. Việc điều chỉnh này nhằm giúp bản dịch được chính xác và rõ nghĩa hơn. Tuy vậy, mọi sai sót của Bản dịch do nhóm biên dịch xin chịu trách nhiệm. Nhóm biên dịch rất mong nhận được sự góp ý của người đọc. Mọi góp ý về bản dịch có thể gửi về cho Trung tâm Thông tin – Thư viện Trường Đại học Nguyễn Tất Thành.

Nhóm biên dịch

Preface for Introductory Business Statistics 2e

Welcome to **Introductory Business Statistics 2e**, an OpenStax resource. This textbook was written to increase student access to high-quality learning materials, maintaining highest standards of academic rigor at little to no cost.

About OpenStax

OpenStax is a nonprofit based at Rice University, and it's our mission to improve student access to education. Our first openly licensed college textbook was published in 2012, and our library has since scaled to over 25 books for college and AP® courses used by hundreds of thousands of students. OpenStax Tutor, our low-cost personalized learning tool, is being used in college courses throughout the country. Through our partnerships with philanthropic foundations and our alliance with other educational resource organizations, OpenStax is breaking down the most common barriers to learning and empowering students and instructors to succeed.

About OpenStax resources

Customization

Introductory Business Statistics 2e is licensed under a Creative Commons Attribution 4.0 International (CC BY) license, which means that you can distribute, remix, and build upon the content, as long as you provide attribution to OpenStax and its content contributors.

Because our books are openly licensed, you are free to use the entire book or pick and choose the sections that are most relevant to the needs of your course. Feel free to remix the content by assigning your students certain chapters and sections in your syllabus, in the order that you prefer. You can even provide a direct link in your syllabus to the sections in the web view of your book.

Instructors also have the option of creating a customized version of their OpenStax book. The custom version can be made available to students in low-cost print or digital form through their campus bookstore. Visit the Instructor Resources section of your book page on OpenStax.org for more information.

Art Attribution

In Introductory Business Statistics 2e, most art contains attribution to its title, creator or rights holder, host platform, and license within the caption. Because the art is openly licensed, anyone may reuse the art as long as they provide the same attribution to its original source.

To maximize readability and content flow, some art does not include attribution in the text. If you reuse illustrations, graphs, or charts from this text that do not have attribution provided, use the following attribution: Copyright Rice University, OpenStax, under CC BY 4.0 license.

Errata

All OpenStax textbooks undergo a rigorous review process. However, like any professional-grade textbook, errors sometimes occur. Since our books are web based, we can make updates periodically when deemed pedagogically necessary. If you have a correction to suggest, submit it through the link on your book page on OpenStax.org. Subject matter experts review all errata suggestions. OpenStax is committed to remaining transparent about all updates, so you will also find a list of past errata changes on your book page on OpenStax.org.

Format

You can access this textbook for free in web view or PDF through OpenStax.org, and for a low cost in print.

About Introductory Business Statistics 2e

Introductory Business Statistics 2e is designed to meet the scope and sequence requirements of the one-semester statistics course for business, economics, and related majors. Core statistical concepts and skills have been augmented with practical business examples, scenarios, and exercises. The result is a meaningful understanding of the discipline which will serve students in their business careers and real-world experiences.

Coverage and scope

Introductory Business Statistics 2e began as a customized version of OpenStax Introductory Statistics by Barbara Illowsky and Susan Dean. Statistics faculty at The University of Oklahoma have used the business statistics adaptation for several years, and the author has continually refined it based on student success and faculty feedback.

The book is structured in a similar manner to most traditional statistics textbooks. The most significant topical changes occur in the latter chapters on regression analysis. Discrete probability density functions have been reordered to provide a logical progression from simple counting formulas to more complex continuous distributions. Many additional homework assignments have been added, as well as new, more mathematical examples.

Introductory Business Statistics 2e places a significant emphasis on the development and practical application of formulas so that students have a deeper understanding of their interpretation and application of data. To achieve this unique approach, the author included a wealth of additional material and purposely de-emphasized the use of the scientific calculator. Specific changes regarding formula use include:

- Expanded discussions of the combinatorial formulas, factorials, and sigma notation
- Adjustments to explanations of the acceptance/rejection rule for hypothesis testing, as well as a focus on terminology regarding confidence intervals
- Deep reliance on statistical tables for the process of finding probabilities (which would not be required if probabilities relied on scientific calculators)

- Continual and emphasized links to the Central Limit Theorem throughout the book; Introductory Business Statistics 2e consistently links each test statistic back to this fundamental theorem in inferential statistics

Another fundamental focus of the book is the link between statistical inference and the scientific method. Business and economics models are fundamentally grounded in assumed relationships of cause and effect. They are developed to both test hypotheses and to predict from such models. This comes from the belief that statistics is the gatekeeper that allows some theories to remain and others to be cast aside for a new perspective of the world around us. This philosophical view is presented in detail throughout and informs the method of presenting the regression model, in particular.

The correlation and regression chapter includes confidence intervals for predictions, alternative mathematical forms to allow for testing categorical variables, and the presentation of the multiple regression model.

Pedagogical features

- Examples are placed strategically throughout the text to show students the step-by-step process of interpreting and solving statistical problems. To keep the text relevant for students, the examples are drawn from a broad spectrum

of practical topics; these include examples about college life and learning, health and medicine, retail and business, and sports and entertainment.

- Practice, Homework, and Bringing It Together give the students problems at various degrees of difficulty while also including real-world scenarios to engage students.
- Try It practice problems immediately follow many examples and give students the opportunity to practice as they read the text.

Changes to the Second Edition

The revision of Introductory Business Statistics has been undertaken with the support of many faculty adopters and student users. Most of the edits are focused on currency, clarity, accuracy, and belonging. The author made a number of additions in order to provide more support and practice opportunities for students; these include conceptual explanations, introductory text, new Try It exercises, and more detailed example solution steps. Also, dozens of examples and references have been changed in order to align with contemporary contexts and uses. Accuracy checking and errata resolutions resulted in corrections and clarifications, both in the main text and in the answers/solutions. Finally, the second edition includes several hundred edits related to gender, race, ethnicity, age, academic status, and societal issues, designed to present the most informed and inclusive material.

OpenStax only undertakes revisions when pedagogically necessary, and we work to ensure that the text maintains its organization in order to ease transition for instructors. Many of the above changes have been made within problems and examples. However, in nearly all cases, the problems retain their original numbering, so that instructors and online homework providers can move to the new edition with minimal impact on their

assignment structures. (This will not be the case for Try It exercises in many chapters, due to the addition of new exercises.) A detailed transition guide is available on the instructor resource page for this textbook.

Answers to Questions in the Book

Answers to Examples are provided directly under the question. Answers to Try it Questions are provided at the Student Resources page. Answers to an assortment of Practice, Homework, and Bringing it Together questions are available to students in the solutions section at the end of each chapter. All other answers to these questions are provided only to instructors in the Instructor Answer Guide via the Instructor Resources page.

Additional resources

Student and instructor resources

We've compiled additional resources for both students and instructors, including Getting Started Guides, an instructor solution manual, and PowerPoint slides. Instructor resources require a verified instructor account, which you can apply for when you log in or create your account on OpenStax.org. Take advantage of these resources to supplement your OpenStax book.

Community Hubs

OpenStax partners with the Institute for the Study of Knowledge Management in Education (ISKME) to offer Community Hubs on OER Commons – a platform for instructors to share community-created resources that support OpenStax books, free of charge. Through our Community Hubs, instructors can upload their own materials or download resources to use in their own courses, including additional ancillaries, teaching material, multimedia, and relevant course content. We encourage instructors to join the hubs for the subjects most relevant to your teaching and research as an opportunity both to enrich your courses and to engage with other faculty.

To reach the Community Hubs, visit <https://www.oercommons.org/hubs/OpenStax>.

Technology partners

As allies in making high-quality learning materials accessible, our technology partners offer optional low-cost tools that are integrated with OpenStax books. To access the technology options for your text, visit your book page on OpenStax.org.

About the authors

Senior contributing authors

Alexander Holmes, The University of Oklahoma

Barbara Illowsky, DeAnza College

Susan Dean, DeAnza College

Contributing authors

Kevin Hadley, Analyst, Federal Reserve Bank of Kansas City

Reviewers

Birgit Aquilonius, West Valley College
Charles Ashbacher, Upper Iowa University - Cedar Rapids
Abraham Biggs, Broward Community College
Daniel Birmajer, Nazareth College
Roberta Bloom, De Anza College
Bryan Blount, Kentucky Wesleyan College
Ernest Bonat, Portland Community College
Sarah Boslaugh, Kennesaw State University
David Bosworth, Hutchinson Community College
Sheri Boyd, Rollins College
George Bratton, University of Central Arkansas
Franny Chan, Mt. San Antonio College
Jing Chang, College of Saint Mary
Laurel Chiappetta, University of Pittsburgh
Lenore Desilets, De Anza College
Matthew Einsohn, Prescott College
Ann Flanigan, Kapiolani Community College
David French, Tidewater Community College
Mo Geraghty, De Anza College
Larry Green, Lake Tahoe Community College
Michael Greenwich, College of Southern Nevada
Inna Grushko, De Anza College
Valier Hauber, De Anza College
Janice Hector, De Anza College
Jim Helmreich, Marist College
Robert Henderson, Stephen F. Austin State University
Mel Jacobsen, Snow College
Mary Jo Kane, De Anza College
John Kagochi, University of Houston - Victoria
Lynette Kenyon, Collin County Community College
Charles Klein, De Anza College
Alexander Kolovos
Sheldon Lee, Viterbo University
Sara Lenhart, Christopher Newport University
Wendy Lightheart, Lane Community College
Vladimir Logvenenko, De Anza College
Jim Lucas, De Anza College
Suman Majumdar, University of Connecticut
Lisa Markus, De Anza College
Miriam Masullo, SUNY Purchase
Diane Mathios, De Anza College
Robert McDevitt, Germanna Community College
John Migliaccio, Fordham University
Mark Mills, Central College
Cindy Moss, Skyline College
Nydia Nelson, St. Petersburg College
Benjamin Ngwudike, Jackson State University
Jonathan Oaks, Macomb Community College

Carol Olmstead, De Anza College
Barbara A. Osyk, The University of Akron
Adam Pennell, Greensboro College
Kathy Plum, De Anza College
Lisa Rosenberg, Elon University
Sudipta Roy, Kankakee Community College
Javier Rueda, De Anza College
Yvonne Sandoval, Pima Community College
Rupinder Sekhon, De Anza College
Travis Short, St. Petersburg College
Frank Snow, De Anza College
Abdulhamid Sukar, Cameron University
Jeffery Taub, Maine Maritime Academy
Mary Teegarden, San Diego Mesa College
John Thomas, College of Lake County
Philip J. Verrecchia, York College of Pennsylvania
Dennis Walsh, Middle Tennessee State University
Cheryl Wartman, University of Prince Edward Island
Carol Weideman, St. Petersburg College
Kyle S. Wells, Dixie State University
Andrew Wiesner, Pennsylvania State University

1 LẤY MẪU VÀ DỮ LIỆU



Hình 1.1 Chúng ta tiếp xúc với thống kê trong đời sống hằng ngày nhiều hơn chúng ta vẫn nghĩ, chúng đến từ vô vàn nguồn khác nhau—chẳng hạn như các bản tin. (nguồn: chỉnh sửa từ tác phẩm “New office” của Phil Whitehouse, Flickr, CC BY 2.0)

Giới thiệu

Có lẽ bạn đang tự hỏi: “Tôi sẽ cần dùng thống kê vào việc gì, lúc nào và ở đâu?” Chỉ cần đọc báo, xem truyền hình hay lướt Internet, bạn sẽ thấy thông tin thống kê xuất hiện khắp nơi: từ tội phạm, thể thao, giáo dục, chính trị đến bất động sản. Thông thường, khi đọc một bài báo hoặc xem bản tin thời sự, bạn được cung cấp các con số dựa trên những mẫu dữ liệu. Từ đó, bạn phải tự đánh giá mức độ chính xác của một tuyên bố, khẳng định, hay “sự thật”. Các phương pháp thống kê giúp bạn đưa ra những “phán đoán có căn cứ tốt nhất”.

Chắc chắn trong cuộc sống bạn sẽ thường xuyên gặp dữ liệu thống kê dưới nhiều hình thức; nên việc nắm vững một số kỹ thuật phân tích là rất cần thiết. Hãy nghĩ đến việc mua nhà, quản lý ngân sách cá nhân hay ngành nghề bạn lựa chọn: kinh tế, kinh doanh, tâm lý học, giáo dục, sinh học, luật, khoa học máy tính, khoa học hình sự, phát triển trẻ nhỏ...—tất cả đều yêu cầu ít nhất một khóa học thống kê.

Chương này sẽ giới thiệu các khái niệm và thuật ngữ cơ bản về xác suất và thống kê. Bạn sẽ nhanh chóng nhận ra rằng hai lĩnh vực này luôn đi kèm nhau. Bạn cũng sẽ học cách thu thập dữ liệu và phân biệt giữa dữ liệu “tốt” và dữ liệu “kém chất lượng”.

1.1 Định nghĩa thống kê, xác suất và các thuật ngữ chính

Thống kê học là ngành khoa học nghiên cứu việc thu thập, phân tích, diễn giải và trình bày **dữ liệu**. Trong đời sống hằng ngày, chúng ta thường xuyên quan sát và sử dụng dữ liệu.

Trong Giáo trình này, bạn sẽ học cách tổ chức và tóm tắt dữ liệu, hoạt động này gọi là **thống kê mô tả** (descriptive statistics). Bạn sẽ tìm hiểu về hai cách phổ biến để tóm tắt dữ liệu là bằng cách lập biểu đồ và bằng cách sử dụng các con số (ví dụ, tìm giá trị trung bình). Bạn sẽ nghiên cứu về **xác suất** (probability) và **phân phối xác suất** (probability distributions) trước khi bạn áp dụng các phương pháp hình thức để rút ra kết luận từ dữ liệu “tốt”. Những phương pháp hình thức này được gọi là **thống kê suy luận** (inferential statistics). Thống kê suy luận sử dụng xác suất để xác định mức độ tin cậy đối với các kết luận của chúng ta.

Diễn giải dữ liệu hiệu quả (suy luận) cần dựa trên quy trình thu thập dữ liệu hợp lý và phân tích dữ liệu một cách thấu đáo. Bạn sẽ gặp khá nhiều công thức toán học dùng để diễn giải dữ liệu. Tuy nhiên, mục tiêu của thống kê không phải là thực hiện vô số phép tính với những công thức này, mà là đạt được sự hiểu biết về dữ liệu của bạn. Việc tính toán có thể được thực hiện bằng máy tính cầm tay hoặc phần mềm máy tính, nhưng sự hiểu biết phải đến từ chính bạn. Nếu bạn có thể nắm vững các nguyên lý cơ bản của thống kê, bạn sẽ trở nên tự tin hơn khi đưa ra các quyết định trong cuộc sống.

Xác suất

Xác suất là một công cụ toán học dùng để nghiên cứu sự ngẫu nhiên. Nó liên quan đến khả năng (cơ hội) xảy ra của một sự kiện. Ví dụ, nếu bạn tung một đồng xu **công bằng** bốn lần (đồng xu công bằng là đồng xu có tỷ lệ xấp, ngửa bằng nhau), kết quả có thể không phải lúc nào cũng là hai mặt ngửa và hai mặt sấp. Tuy nhiên, nếu bạn tung đồng xu đó 4.000 lần, kết quả sẽ gần như chia đều giữa mặt ngửa và mặt sấp. Xác suất lý thuyết kỳ vọng cho mặt ngửa trong mỗi lần tung là 0,5. Mặc dù kết quả của một số lần thử ban đầu có thể không chắc chắn, nhưng khi số lần lặp lại tăng lên, các kết quả sẽ tuân theo một mẫu kết quả có quy luật. Nhà thống kê người Anh Karl Pearson — người đã tung đồng xu 24.000 lần và thu được 12.012 lần mặt ngửa; một trong các tác giả của quyển sách này cũng đã thử tung đồng xu 2.000 lần và kết quả thu được là 996 lần mặt ngửa. Phân số $\frac{996}{2000}$ bằng 0,498, rất gần với giá trị kỳ vọng là 0,5.

Lý thuyết xác suất khởi nguồn từ việc nghiên cứu các trò chơi may rủi, chẳng hạn như poker. Các dự đoán được biểu diễn dưới dạng xác suất. Để dự đoán khả năng xảy ra của một trận động đất, mưa, hoặc khả năng bạn sẽ đạt điểm A trong khóa học này, chúng ta sử dụng xác suất. Trong y học, các bác sĩ dùng xác suất để đánh giá khả năng một loại vắc-xin có thể gây ra chính căn bệnh mà nó được thiết kế để phòng ngừa. Một nhà môi giới chứng khoán sử dụng xác suất để ước lượng tỷ suất sinh lợi từ các khoản đầu tư của khách hàng. Thậm chí, bạn có thể dùng xác suất để cân nhắc xem có nên mua vé số hay không. Trong quá trình học thống kê, bạn sẽ vận dụng sức mạnh của toán học thông qua các phép tính xác suất để phân tích và diễn giải dữ liệu của mình.

Các thuật ngữ chính

Trong thống kê, chúng ta thường hướng đến việc nghiên cứu một **quần thể**¹ (population). Bạn có thể hiểu quần thể là một tập hợp các cá thể, sự vật hoặc đối tượng nằm trong phạm vi nghiên cứu. Để nghiên cứu quần thể, chúng ta sẽ chọn ra một **mẫu** (sample). Ý tưởng của việc **lấy mẫu** (sampling) là chọn một phần (hoặc một tập con) từ quần thể lớn hơn và

¹ Còn gọi là đám đông.

ngiên cứu phần đó (tức là mẫu) nhằm thu thập thông tin đại diện cho toàn bộ quần thể. Dữ liệu chính là kết quả của việc lấy mẫu từ quần thể.

Vì việc kiểm tra toàn bộ quần thể thường tiêu tốn rất nhiều thời gian và chi phí, nên lấy mẫu trở thành một kỹ thuật cực kỳ thiết thực. Ví dụ, nếu bạn muốn tính điểm trung bình chung (overall grade point average) của trường mình, thì khi chọn một mẫu những sinh viên đang theo học tại trường sẽ hợp lý hơn. Dữ liệu thu thập từ mẫu đó chính là điểm trung bình chung của các sinh viên trong mẫu. Ví dụ, trong các cuộc bầu cử tổng thống tại Mỹ, các cuộc thăm dò ý kiến (opinion poll) thường lấy mẫu từ 1.000 đến 2.000 người. Các cuộc thăm dò ý kiến này nhằm đại diện cho quan điểm của toàn thể dân số cả nước Mỹ. Tương tự, các nhà sản xuất nước giải khát có ga đóng lon sẽ lấy mẫu để kiểm tra xem một lon 330 ml có thực sự chứa đủ 330 ml nước hay không.

Từ dữ liệu mẫu, chúng ta có thể tính toán một thông số thống kê. **Thông số thống kê** (statistic) là một con số đại diện cho một đặc tính của mẫu. Ví dụ, nếu chúng ta coi một lớp toán là mẫu đại diện cho quần thể tất cả các lớp toán, thì số điểm bình quân mà sinh viên trong lớp đó đạt được vào cuối kỳ là một ví dụ về thông số thống kê. Thông số thống kê này đóng vai trò là ước lượng điểm (point estimate) cho một tham số của quần thể, trong trường hợp này là bình quân số học (mean). **Tham số** (parameter) là một đặc tính số học của toàn bộ quần thể và có thể được ước lượng thông qua thông số thống kê. Vì chúng ta coi tất cả các lớp toán là quần thể, nên số điểm bình quân mà sinh viên đạt được trong tất cả các lớp toán chính là một ví dụ về tham số.

Một trong những mối quan tâm chính trong lĩnh vực thống kê là mức độ chính xác khi một thông số thống kê ước lượng một tham số. Độ chính xác này phụ thuộc vào mức độ mà mẫu phản ánh đúng đặc điểm của quần thể. Để mẫu có thể được xem là một **mẫu đại diện** (representative sample), nó phải chứa các đặc điểm tiêu biểu của quần thể. Trong thống kê suy luận, chúng ta quan tâm đến cả thông số thống kê mẫu và tham số quần thể. Ở các chương sau, chúng ta sẽ sử dụng thông số thống kê mẫu để kiểm định tính hợp lệ của tham số đã được thiết lập cho quần thể.

Một **biến** (variable), hay biến ngẫu nhiên (random variable), thường được ký hiệu bằng các chữ cái in hoa như X và Y , là một đặc tính hoặc phép đo có thể xác định được cho từng phần tử của quần thể. Biến có thể là **biến số trị** (numerical variable) hoặc **biến phân loại** (categorical variable). Biến số trị nhận các giá trị có đơn vị giống nhau như trọng lượng tính bằng kilogram và thời gian tính bằng giờ. Biến phân loại đặt cá thể hoặc vật thể vào các nhóm nhất định. Nếu đặt X là số điểm mà một sinh viên toán đạt được vào cuối kỳ, thì X là một biến số trị. Nếu đặt Y là đảng phái chính trị của một cá nhân, thì các giá trị của Y có thể là Cộng hòa, Dân chủ hoặc Độc lập. Khi đó, Y là một biến phân loại. Chúng ta có thể thực hiện các phép toán với các giá trị của X (chẳng hạn tính bình quân số điểm đạt được). Tuy nhiên, chúng ta không thể thực hiện các phép toán với các giá trị của Y (tính bình quân đảng phái chính trị không có ý nghĩa).

Dữ liệu (data, số ít datum) là các giá trị thực tế của biến. Chúng có thể là số hoặc từ.

Hai thuật ngữ thường xuyên được sử dụng trong thống kê là **số bình quân** (mean) và **tỷ lệ** (proportion). Nếu bạn thực hiện ba bài kiểm tra trong lớp toán và đạt điểm 86, 75 và 92, bạn sẽ tính điểm bình quân bằng cách cộng ba điểm số đó lại và chia cho ba (điểm bình quân của bạn sẽ là 84,3 làm tròn đến một chữ số thập phân). Nếu trong lớp toán của bạn

có 40 sinh viên, trong đó có 22 nam và 18 nữ, thì tỷ lệ sinh viên nam là $\frac{22}{40}$ và tỷ lệ sinh viên nữ là $\frac{18}{40}$. Số bình quân và tỷ lệ sẽ được trình bày chi tiết hơn ở các chương tiếp theo.

LƯU Ý

Các từ "**số bình quân** (mean)" và "**trung bình** (average)" thường được sử dụng thay thế cho nhau. Việc thay thế này là một thực tiễn phổ biến. Về mặt thuật ngữ, "bình quân số học" (arithmetic mean) là tham số, còn "trung bình" (average) là thông số thống kê, cả hai đều đề cập đến tính hướng tâm của dữ liệu. Tuy nhiên, trong thực tế — đặc biệt là đối với những người không chuyên về thống kê — "trung bình" thường được chấp nhận để thay thế cho "bình quân số học".

Ví dụ 1.1

Vấn đề

Xác định các thuật ngữ chính đề cập đến điều gì trong nghiên cứu sau. Chúng ta muốn biết số tiền trung bình (số bình quân) mà sinh viên năm nhất tại Trường Cao đẳng ABC chi tiêu cho đồ dùng học tập (không bao gồm sách). Chúng ta đã khảo sát ngẫu nhiên 100 sinh viên năm nhất của trường. Ba trong số những sinh viên này đã chi tiêu lần lượt 150 đô la, 200 đô la và 225 đô la.

Lời giải

Quần thể (population) là tất cả sinh viên năm nhất đang theo học tại Trường Cao đẳng ABC trong học kỳ này.

Mẫu (sample) có thể là tất cả sinh viên đăng ký học trong một lớp thống kê cơ bản tại Trường Cao đẳng ABC (mặc dù mẫu này có thể không đại diện cho toàn bộ quần thể).

Tham số (parameter) là số tiền trung bình (số bình quân) đã chi tiêu (không bao gồm tiền sách) trong học kỳ này của sinh viên năm nhất tại trường Cao đẳng ABC: số bình quân của quần thể.

Thông số thống kê (statistic) là số tiền trung bình (số bình quân) đã chi tiêu (không bao gồm sách) của sinh viên năm nhất trong mẫu.

Biến (variable) có thể là số tiền chi tiêu (không bao gồm sách) của một sinh viên năm nhất. Giả sử X = số tiền chi tiêu (không bao gồm sách) của một sinh viên năm nhất tại Trường Cao đẳng ABC.

Dữ liệu (data) là số tiền đô la mà sinh viên năm nhất chi tiêu. Ví dụ về dữ liệu là 150 đô la, 200 đô la và 225 đô la.

Bài tập vận dụng 1.1

Xác định các thuật ngữ chính đề cập đến điều gì trong nghiên cứu sau. Chúng ta muốn biết số tiền trung bình (số bình quân) mà các gia đình có con học tại Học viện Knoll chi tiêu cho đồng phục học sinh mỗi năm. Chúng ta đã khảo sát ngẫu nhiên 100 gia đình có

con đang theo học tại trường. Ba trong số các gia đình đã chi tiêu lần lượt là 65 đô la, 75 đô la và 95 đô la.

Ví dụ 1.2

Vấn đề

Xác định các thuật ngữ chính đề cập đến điều gì trong nghiên cứu sau.

Một nghiên cứu đã được tiến hành tại một trường cao đẳng địa phương nhằm phân tích điểm trung bình tích lũy (GPA) của các sinh viên đã tốt nghiệp trong năm vừa qua. Hãy điền chữ cái tương ứng với cụm từ mô tả phù hợp nhất cho từng mục sau:

1. Quần thể ___ 2. Thông số thống kê ___ 3. Tham số ___ 4. Mẫu ___ 5. Biến ___ 6. Dữ liệu ___
- a. tất cả sinh viên đã học tại trường cao đẳng năm vừa qua.
- b. điểm trung bình tích lũy (GPA) của một sinh viên đã tốt nghiệp trường cao đẳng học năm vừa qua
- c. 3,65; 2,80; 1,50; 3,90.
- d. một nhóm sinh viên đã tốt nghiệp trường cao đẳng trong năm vừa qua, được chọn ngẫu nhiên
- e. điểm trung bình tích lũy GPA của các sinh viên đã tốt nghiệp trường cao đẳng trong năm vừa qua
- f. tất cả sinh viên đã tốt nghiệp trường cao đẳng trong năm vừa qua
- g. điểm trung bình tích lũy GPA của các sinh viên trong nghiên cứu đã tốt nghiệp trường cao đẳng trong năm vừa qua

Lời giải

1. f; 2. g; 3. e; 4. d; 5. b; 6. c

Bài tập vận dụng 1.2

Xác định những thuật ngữ chính được đề cập trong nghiên cứu sau.

Một cuộc khảo sát đối với các vận động viên tại một trường đại học đã được thực hiện nhằm nghiên cứu chiều cao của họ (tính bằng mét). Hãy điền chữ cái tương ứng với cụm từ mô tả phù hợp nhất cho từng mục dưới đây:

1. Quần thể (population) ___
 2. Thông số thống kê (statistic) ___
 3. Tham số (parameter) ___
 4. Mẫu (sample) ___
 5. Biến (variable) ___
 6. Dữ liệu (data) ___
- a. chiều cao trung bình của các vận động viên trong trường đại học
 - b. chiều cao trung bình của các vận động viên trong cuộc khảo sát
 - c. tất cả các vận động viên trong trường đại học

- d. tất cả sinh viên trong trường đại học
- e. chiều cao của một vận động viên
- f. một nhóm vận động viên được chọn ngẫu nhiên
- g. 1,82; 1,76; 1,69; 1,93.

Ví dụ 1.3*Vấn đề*

Xác định các thuật ngữ chính đề cập đến điều gì trong nghiên cứu sau.

Là một phần của nghiên cứu được thiết kế để kiểm tra mức độ an toàn của ô tô điện, Ủy ban An toàn Giao thông Quốc gia đã thu thập và xem xét dữ liệu về tác động của một vụ va chạm ô tô đối với các người nộm (dummies) trong thử nghiệm. Tiêu chí được sử dụng như sau:

Bảng 1.1

Tốc độ va chạm	Vị trí của “tài xế” (người nộm)
35 dặm/giờ	Ghế trước

Các ô tô có người nộm ở ghế trước được cho va chạm vào tường ở tốc độ 35 dặm/giờ. Mục tiêu là xác định tỷ lệ người nộm ở vị trí tài xế có thể bị chấn thương ở đầu, giả định rằng họ là người thật. Một mẫu ngẫu nhiên đơn giản gồm 75 chiếc xe đã được chọn.

Lời giải

Quần thể là tất cả các ô tô có người nộm ở ghế trước.

Mẫu bao gồm 75 chiếc xe được chọn bằng phương pháp lấy mẫu ngẫu nhiên đơn giản.

Tham số là tỷ lệ người nộm ở vị trí tài xế (nếu là người thật) bị chấn thương đầu trong quần thể.

Thông số thống kê là tỷ lệ người nộm ở vị trí tài xế (nếu là người thật) bị chấn thương đầu trong mẫu.

Biến X = Liệu một người nộm (nếu là người thật) có bị chấn thương đầu hay không.

Dữ liệu: Có (bị chấn thương đầu) hoặc Không (không bị).

Bài tập vận dụng 1.3

Xác định các thuật ngữ chính đề cập đến điều gì trong nghiên cứu sau.

Một cuộc khảo sát được tiến hành nhằm kiểm tra thời gian sạc pin của điện thoại di động từ 50% đến 100%. Các tiêu chí được sử dụng để thu thập dữ liệu như sau:

Bảng 1.2

Công suất sạc	Loại điện thoại di động
30W	Android

Chúng ta muốn biết tỷ lệ điện thoại Android được sạc đầy 100% trong vòng 30 phút. Một mẫu ngẫu nhiên đơn giản gồm 200 điện thoại đã được chọn.

Ví dụ 1.4*Vấn đề*

Xác định các thuật ngữ chính đề cập đến điều gì trong nghiên cứu sau.

Một công ty bảo hiểm muốn xác định tỷ lệ các bác sĩ đã từng liên quan đến một hoặc nhiều vụ kiện về sai sót y tế. Công ty đã chọn ngẫu nhiên 500 bác sĩ từ một danh bạ chuyên môn và xác định số lượng bác sĩ trong mẫu đã từng liên quan đến một vụ kiện sai sót y tế.

Lời giải

Quần thể là tất cả các bác sĩ được liệt kê trong danh bạ chuyên môn.

Tham số là tỷ lệ các bác sĩ đã từng liên quan đến một hoặc nhiều vụ kiện sai sót y tế trong quần thể.

Mẫu là 500 bác sĩ được chọn ngẫu nhiên từ danh bạ chuyên môn.

Thông số thống kê là tỷ lệ bác sĩ trong mẫu đã từng liên quan đến một hoặc nhiều vụ kiện sai sót y tế.

Biến X = liệu một bác sĩ có liên quan đến một vụ kiện sai sót y tế hay không.

Dữ liệu: Có (đã từng liên quan đến một hoặc nhiều vụ kiện) hoặc Không (chưa từng liên quan).

Bài tập vận dụng 1.4

Xác định các thuật ngữ chính đề cập đến điều gì trong nghiên cứu sau.

Một nghiên cứu sẽ được thực hiện bởi một hãng thông tấn để tìm ra tỷ lệ tất cả các tài xế xe tải không có điểm phạt nào trên giấy phép của họ. Cơ quan này chọn ngẫu nhiên 1.000 tài xế xe tải từ danh bạ tài xế và xác định số lượng tài xế xe tải trong mẫu không có điểm phạt nào trên giấy phép.

Bài tập nhóm

Thực hiện bài tập sau theo nhóm tối đa bốn người. Hãy xác định quần thể (population), mẫu (sample), tham số (parameter), thông số thống kê (statistic), biến (variable) và dữ liệu (data) cho nghiên cứu sau: Bạn muốn xác định số lượng trung bình (bình quân) các

ly sữa mà sinh viên đại học uống mỗi ngày. Giả sử ngày hôm qua, trong giờ học tiếng Anh, bạn đã hỏi năm sinh viên về số ly sữa họ đã uống vào ngày hôm trước. Các câu trả lời lần lượt là 1, 0, 1, 3 và 4 ly sữa.

1.2 Dữ liệu, lấy mẫu và biến thiên trong dữ liệu và lấy mẫu

Dữ liệu có thể bắt nguồn từ quần thể hoặc từ mẫu. Các chữ cái viết thường như x hoặc y thường được dùng để biểu diễn các giá trị dữ liệu. Phần lớn dữ liệu có thể được phân loại vào các nhóm sau:

- Định tính (Qualitative)
- Định lượng (Quantitative)

Dữ liệu định tính là kết quả của việc phân loại hoặc mô tả các thuộc tính của một quần thể. Dữ liệu định tính cũng thường được gọi là dữ liệu phân loại. Màu tóc, nhóm máu, nhóm dân tộc, loại xe mà một người sử dụng, và tên con đường nơi người đó sinh sống là các ví dụ điển hình của dữ liệu định tính (dữ liệu phân loại). Dữ liệu định tính (dữ liệu phân loại) thường được mô tả bằng từ ngữ hoặc chữ cái. Ví dụ, màu tóc có thể là đen, nâu sẫm, nâu nhạt, vàng hoe, xám hoặc đỏ. Nhóm máu có thể là AB+, O- hoặc B+. Các nhà nghiên cứu thường ưu tiên sử dụng dữ liệu định lượng thay vì dữ liệu định tính (phân loại) vì dữ liệu định lượng dễ dàng áp dụng các phương pháp phân tích toán học. Chẳng hạn, việc tính giá trị trung bình của màu tóc hoặc nhóm máu là không có ý nghĩa.

Dữ liệu định lượng được thể hiện dưới dạng số. Đây là kết quả của **việc đếm** hoặc **đo lường** các thuộc tính của một quần thể. Số tiền, nhịp tim, trọng lượng cơ thể, số người sống trong một thị trấn, và số sinh viên theo học môn thống kê là những ví dụ tiêu biểu của dữ liệu định lượng. Dữ liệu định lượng có thể là **rời rạc** hoặc **liên tục**.

Tất cả dữ liệu thu được từ việc đếm đều được gọi là **dữ liệu định lượng rời rạc**. Loại dữ liệu này chỉ nhận giá trị số nhất định. Nếu bạn đếm số cuộc gọi nhận được mỗi ngày trong tuần, bạn có thể ghi nhận các giá trị như 0, 1, 2 hoặc 3.

Dữ liệu không chỉ bao gồm các số nguyên mà còn có thể bao gồm phân số, số thập phân, hoặc số vô tỉ được gọi là **dữ liệu định lượng liên tục**. Dữ liệu liên tục thường là kết quả của các phép đo như độ dài, trọng lượng hoặc thời gian. Danh sách độ dài (tính bằng phút) của tất cả các cuộc gọi bạn thực hiện trong một tuần — với các giá trị như 2,4, 7,5 hoặc 11,0 — chính là ví dụ của dữ liệu định lượng liên tục.

Ví dụ 1.5

Mẫu dữ liệu định lượng rời rạc

Dữ liệu thu thập là số lượng sách mà sinh viên mang theo trong ba lô của họ. Bạn lấy mẫu năm sinh viên. Hai sinh viên mang theo ba quyển sách, một sinh viên mang theo bốn quyển, một sinh viên mang theo hai quyển, và một sinh viên mang theo một quyển. Các con số về số sách (ba, bốn, hai và một) chính là dữ liệu định lượng rời rạc.

Bài tập vận dụng 1.5

Dữ liệu là số lượng máy tập trong một phòng gym. Bạn lấy mẫu năm phòng gym. Một phòng có 12 máy, một phòng có 15 máy, một phòng có 10 máy, một phòng có 22 máy, và phòng còn lại có 20 máy. Đây là loại dữ liệu gì?

Ví dụ 1.6**Mẫu dữ liệu định lượng liên tục**

Dữ liệu là trọng lượng của ba lô chứa sách. Bạn lấy mẫu cùng năm sinh viên đó. Trọng lượng (tính bằng pound) của các ba lô là: 6,2, 7, 6,8, 9,1 và 4,3. Lưu ý rằng các ba lô mang cùng ba quyển sách có thể có trọng lượng khác nhau. Trọng lượng là dữ liệu định lượng liên tục.

Bài tập vận dụng 1.6

Dữ liệu là diện tích bãi cỏ (tính bằng mét vuông). Bạn lấy mẫu năm ngôi nhà. Diện tích các bãi cỏ lần lượt là: 144 mét vuông, 160 mét vuông, 190 mét vuông, 180 mét vuông và 210 mét vuông. Đây là loại dữ liệu gì?

Ví dụ 1.7

Bạn đi siêu thị và mua ba lon súp (562,9 ml súp cà chua kem, 416,1 ml súp đậu lăng, và 562,9 ml súp Ý), hai gói hạt (óc chó và đậu phộng), bốn loại rau khác nhau (bông cải xanh, súp lơ trắng, rau chân vịt và cà rốt), và hai món tráng miệng (473,2 ml kem hạt hồ trăn và 946,3 ml bánh quy hạt sô cô la).

Vấn đề

Hãy chỉ ra những tập dữ liệu nào thuộc loại định lượng rời rạc, định lượng liên tục, và định tính (phân loại).

Lời giải

Một lời giải khả thi :

- Ba lon súp, hai gói hạt, bốn loại rau và hai món tráng miệng là dữ liệu định lượng rời rạc vì chúng là kết quả của việc đếm.
- Trọng lượng của các lon súp (562,9 ml, 416,1 ml, 562,9 ml) là dữ liệu định lượng liên tục vì trọng lượng được đo lường với độ chính xác cao nhất có thể.
- Loại súp, loại hạt, loại rau và loại món tráng miệng là dữ liệu định tính (phân loại) vì chúng được phân theo danh mục.

Hãy thử xác định thêm các tập dữ liệu khác có thể có trong ví dụ này.

Hãy thử 1.7

Danh sách sau là các vật tư được một quản lý mua hàng trong một công ty đặt mua:

- Hai loại đinh (đinh hộp 2 kg, đinh lợp mái 3 kg)
- Một loại dầu (dầu máy 4 lít)
- Bốn loại ốc vít (ốc gỗ 3 kg, ốc máy 5 kg, ốc định vị 1 kg, ốc đầu lỗ lục giác 2 kg)

Hãy xác định các tập dữ liệu thuộc loại định lượng rời rạc, định lượng liên tục và định tính.

Ví dụ 1.8

Dữ liệu là màu sắc của ba lô. Bạn tiếp tục lấy mẫu năm sinh viên trước đó. Một sinh viên có ba lô màu đỏ, hai sinh viên có ba lô màu đen, một sinh viên có ba lô màu xanh lá, và một sinh viên có ba lô màu xám. Các màu sắc: đỏ, đen, đen, xanh lá và xám là dữ liệu định tính (phân loại).

Bài tập vận dụng 1.8

Dữ liệu là màu sắc của các ngôi nhà. Bạn lấy mẫu năm ngôi nhà. Màu sắc của các ngôi nhà lần lượt là trắng, vàng, trắng, đỏ và trắng. Đây là loại dữ liệu gì?

LƯU Ý

Bạn có thể thu thập dữ liệu dưới dạng số nhưng trình bày nó theo cách phân loại. Ví dụ, điểm kiểm tra của mỗi sinh viên được ghi nhận trong suốt học kỳ. Đến cuối học kỳ, điểm số được báo cáo dưới dạng xếp loại: A, B, C, D hoặc F.

Ví dụ 1.9*Vấn đề*

Làm việc theo nhóm để xác định đúng loại dữ liệu (định lượng hay định tính). Nếu là dữ liệu định lượng, hãy chỉ rõ đó là dữ liệu liên tục hay rời rạc. Gợi ý: Dữ liệu rời rạc thường bắt đầu bằng cụm từ “số lượng...”

- số lượng đôi giày bạn sở hữu
- loại xe bạn đang sử dụng
- khoảng cách từ nhà bạn đến cửa hàng tạp hóa gần nhất
- số lớp học bạn tham gia trong mỗi năm học.
- loại máy tính cầm tay bạn sử dụng
- trọng lượng của những chú chó tại một trạm cứu hộ động vật

- g. số câu trả lời đúng trong một bài kiểm tra
- h. chỉ số IQ (điều này có thể gây tranh luận)

Lời giải

Các mục a, d, và g là dữ liệu định lượng rời rạc; Các mục c, f, và h là dữ liệu định lượng liên tục; Các mục b và e là dữ liệu định tính (phân loại).

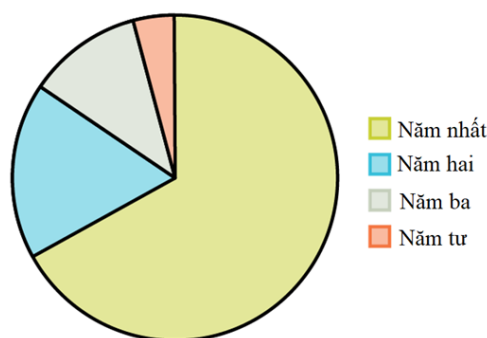
Bài tập vận dụng 1.9

Xác định đúng loại dữ liệu (định lượng hay định tính) cho số lượng xe hơi trong một bãi đậu xe. Nếu là dữ liệu định lượng, hãy chỉ rõ đó là rời rạc hay liên tục.

Ví dụ 1.10

Vấn đề

Một giảng viên thống kê thu thập thông tin về phân loại sinh viên trong lớp của mình theo các nhóm: sinh viên năm nhất, sinh viên năm hai, sinh viên năm ba, hoặc sinh viên năm tư. Dữ liệu thu thập được được tóm tắt trong biểu đồ tròn [Hình 1.2](#). Biểu đồ này hiển thị loại dữ liệu nào?



Hình 1.2 Phân loại sinh viên lớp thống kê

Lời giải

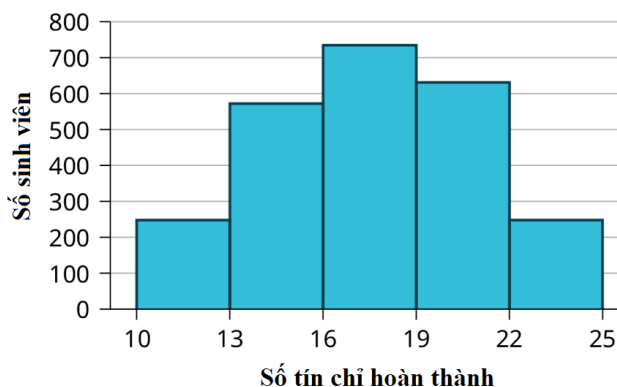
Biểu đồ tròn này thể hiện số lượng sinh viên theo từng năm, đây là **dữ liệu định tính (phân loại)**.

Bài tập vận dụng 1.10

Phòng đào tạo tại Trường Đại học Tiểu bang lưu trữ dữ liệu về số tín chỉ sinh viên hoàn thành mỗi học kỳ. Dữ liệu thu thập sẽ được tóm tắt trong biểu đồ tần suất dạng cột (histogram). Các khoảng giá trị phân nhóm được chia như sau:

- Từ 10 đến dưới 13 tín chỉ

- Từ 13 đến dưới 16 tín chỉ
- Từ 16 đến dưới 19 tín chỉ
- Từ 19 đến dưới 22 tín chỉ
- Từ 22 đến dưới 25 tín chỉ



Hình 1.3 Số tín chỉ hoàn thành của mỗi sinh viên

Biểu đồ này hiển thị loại dữ liệu nào?

Thảo luận về dữ liệu định tính

Dưới đây là các bảng so sánh số lượng sinh viên ghi danh học bán thời gian và toàn thời gian tại Trường Cao đẳng De Anza và Trường Cao đẳng Foothill trong học kỳ mùa xuân gần nhất. Các bảng trình bày **số lượng sinh viên** (tần suất) và **tỷ lệ phần trăm** hoặc **tỷ lệ** (tần suất tương đối). Cột tỷ lệ phần trăm giúp việc so sánh cùng một nhóm sinh viên giữa hai trường trở nên dễ dàng hơn. Việc hiển thị tỷ lệ phần trăm cùng với số lượng tuyệt đối thường rất hữu ích, đặc biệt quan trọng khi so sánh các bộ dữ liệu có tổng số khác nhau — như tổng số lượng ghi danh tại hai trường trong ví dụ này. Ta có thể thấy rằng tỷ lệ phần trăm sinh viên học bán thời gian tại Trường Cao đẳng Foothill cao hơn nhiều so với Trường Cao đẳng De Anza.

Bảng 1.3 Học kỳ mùa xuân gần nhất

Trường Cao đẳng De Anza			Trường Cao đẳng Foothill		
	Số lượng	Tỷ lệ phần trăm		Số lượng	Tỷ lệ phần trăm
Toàn thời gian	9.200	40,9%	Toàn thời gian	4.059	28,6%
Bán thời gian	13.296	59,1%	Bán thời gian	10.124	71,4%
Tổng cộng	22.496	100%	Tổng cộng	14.183	100%

Bảng số liệu là một cách tốt để tổ chức và trình bày dữ liệu. Tuy nhiên, biểu đồ có thể hữu ích hơn trong việc giúp hiểu rõ dữ liệu. Không có quy tắc cứng nhắc nào quy định phải sử

dụng loại biểu đồ nào. Hai loại biểu đồ thường được sử dụng để trình bày dữ liệu định tính (phân loại) là biểu đồ tròn (pie chart) và đồ thị thanh (bar graph).

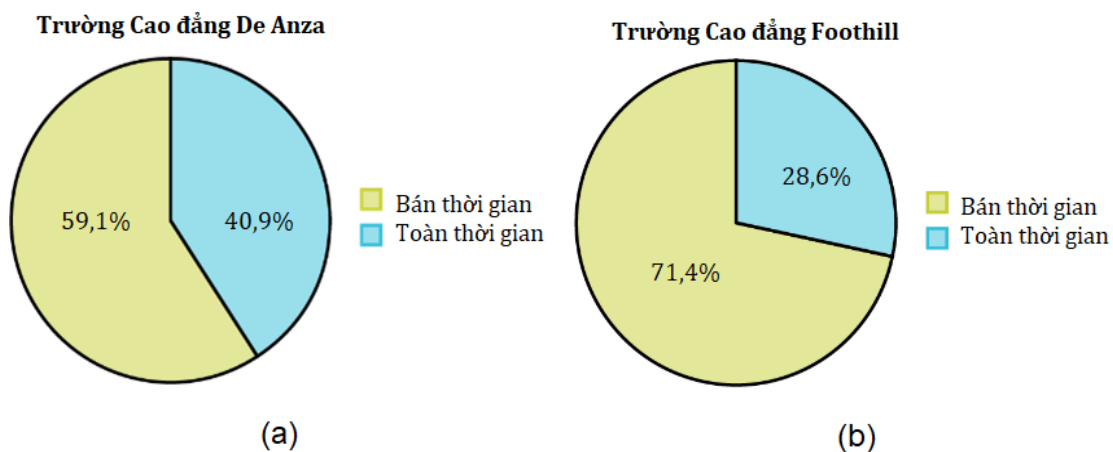
Trong **biểu đồ tròn**, các nhóm dữ liệu được biểu diễn bằng các hình quạt trong vòng tròn và có kích thước tỷ lệ thuận với tỷ lệ phần trăm của từng nhóm trong tổng thể.

Trong **đồ thị thanh**, độ dài của mỗi cột biểu thị số lượng hoặc tỷ lệ phần trăm của từng nhóm dữ liệu. Các cột có thể được sắp xếp theo chiều dọc hoặc chiều ngang.

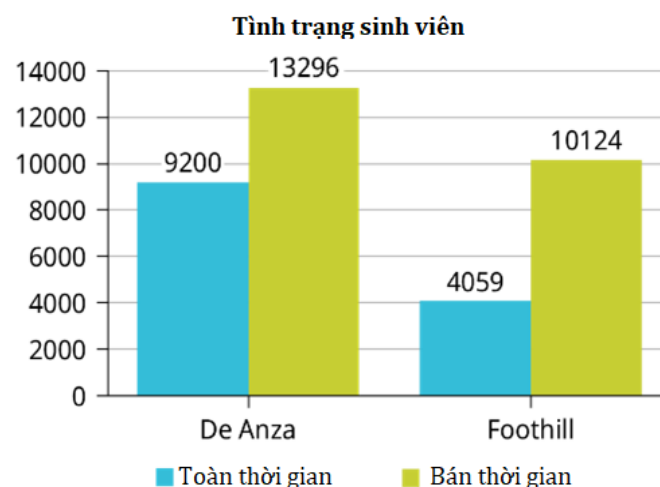
Biểu đồ Pareto bao gồm các cột được sắp xếp theo thứ tự giảm dần theo kích thước nhóm (từ lớn nhất đến nhỏ nhất).

Hãy quan sát Hình 1.4 và Hình 1.5 và xác định biểu đồ nào (biểu đồ tròn hay biểu đồ cột) giúp bạn so sánh dữ liệu dễ dàng hơn.

Chúng ta nên xem xét nhiều loại biểu đồ khác nhau để tìm ra cách trình bày dữ liệu hiệu quả nhất. Thế nào là biểu đồ "tốt nhất" tùy thuộc vào bản chất dữ liệu và bối cảnh sử dụng. Lựa chọn loại biểu đồ cũng phụ thuộc vào mục đích sử dụng dữ liệu.



Hình 1.4



Hình 1.5

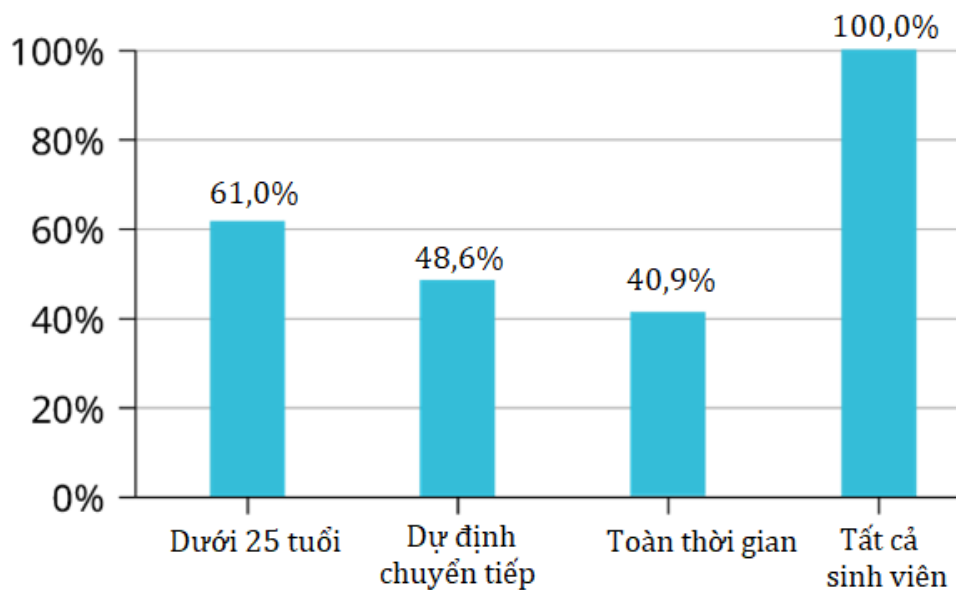
Tổng tỷ lệ phần trăm vượt quá (hoặc thấp hơn) 100%

Đôi khi, tổng các tỷ lệ phần trăm có thể vượt quá 100% (hoặc thấp hơn 100%). Trong đồ thị đi kèm (Hình 1.6), tổng tỷ lệ phần trăm vượt quá 100% là do một sinh viên có thể thuộc nhiều hơn một nhóm dữ liệu.

Đồ thị thanh là công cụ phù hợp để so sánh kích thước tương đối giữa các nhóm trong trường hợp này. Biểu đồ tròn không thể sử dụng được, bởi vì biểu đồ tròn yêu cầu tổng tỷ lệ phần trăm phải đúng bằng 100%. Điều này cũng áp dụng nếu tổng tỷ lệ phần trăm nhỏ hơn 100%.

Bảng 1.4 Trường Cao đẳng De Anza, Học kỳ mùa xuân gần nhất

Đặc điểm/nhóm	Tỷ lệ phần trăm
Sinh viên học toàn thời gian	40,9%
Sinh viên có dự định chuyển tiếp lên cơ sở giáo dục 4 năm	48,6%
Sinh viên dưới 25 tuổi	61,0%
TỔNG CỘNG	150,5%



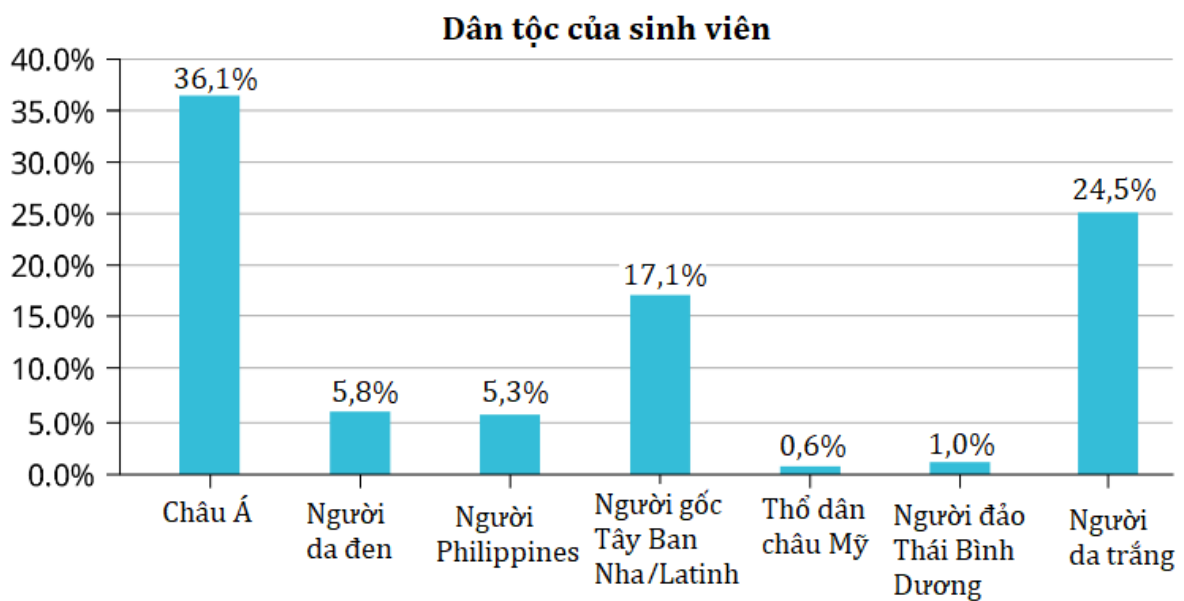
Hình 1.6

Sốt danh mục/Thiếu dữ liệu

Bảng 1.5 trình bày thông tin về dân tộc của sinh viên, nhưng đã thiếu nhóm “Khác/Không xác định”. Nhóm này bao gồm những người cảm thấy họ không thuộc bất kỳ nhóm dân tộc nào được liệt kê hoặc từ chối trả lời. Có thể thấy rằng tổng tần suất không cộng lại bằng tổng số sinh viên. Trong trường hợp như vậy, nên tạo đồ thị thanh, không sử dụng biểu đồ tròn, vì biểu đồ tròn yêu cầu tổng dữ liệu phải đầy đủ (tổng 100%).

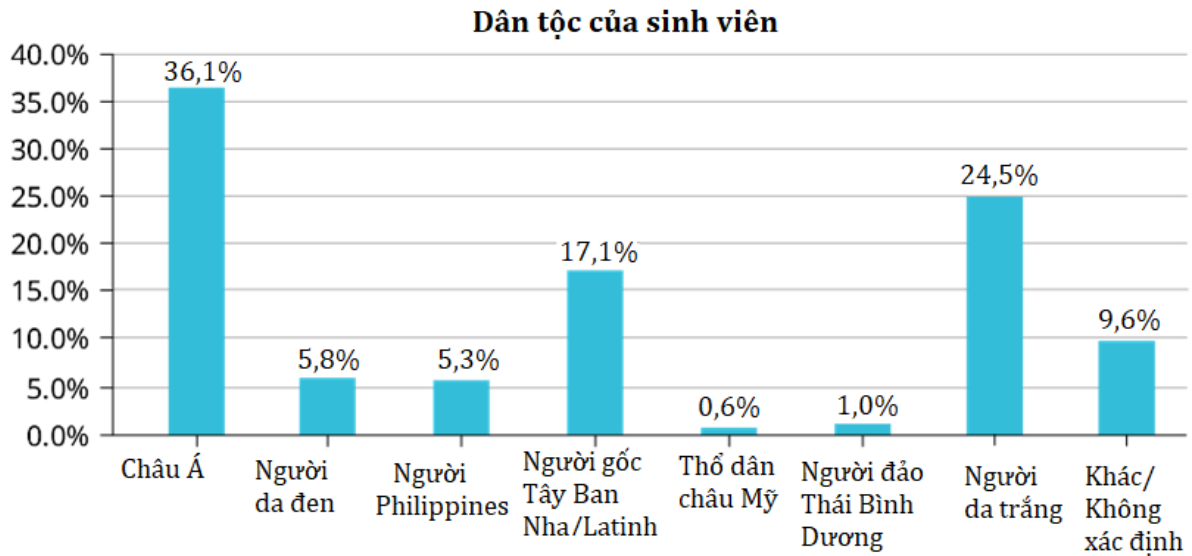
Bảng 1.5 Dân tộc của sinh viên tại Trường Cao đẳng De Anza trong học kỳ mùa thu gần nhất

	Tần suất	Tỷ lệ phần trăm
Châu Á	8.794	36,1%
Người da đen	1.412	5,8%
Người Philippines	1.298	5,3%
Người gốc Tây Ban Nha/Latinh	4.180	17,1%
Thổ dân châu Mỹ	146	0,6%
Người đảo Thái Bình Dương	236	1,0%
Người da trắng	5.978	24,5%
TỔNG CỘNG	22.044 trên tổng số 24.382	90,4% trên tổng số 100%

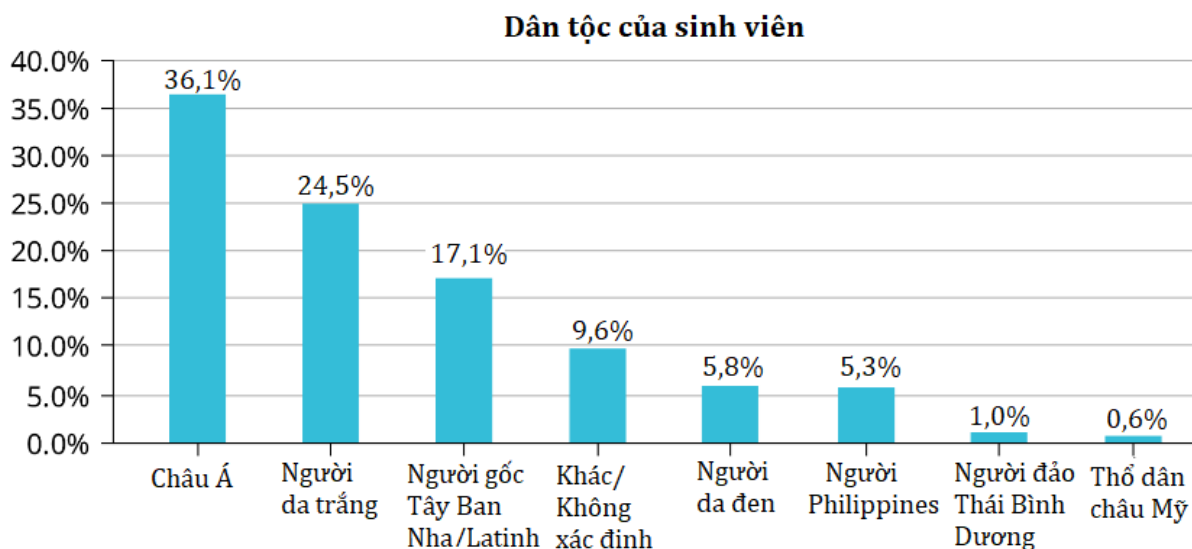
**Hình 1.7**

Đồ thị sau đây (Hình 1.8) giống với đồ thị trước đó (Hình 1.7) nhưng đã bổ sung thêm tỷ lệ phần trăm của nhóm “Khác/Không xác định” (9,6%). Nhóm “Khác/Không xác định” có quy mô khá lớn so với một số nhóm khác (Thổ dân châu Mỹ: 0,6%, Người đảo Thái Bình Dương: 1,0%). Điều này rất quan trọng khi chúng ta xem xét ý nghĩa mà dữ liệu truyền đạt.

Đồ thị thanh trong Hình 1.8 có thể khó hiểu khi quan sát trực quan. Đồ thị trong [Hình 1.9](#) là biểu đồ Pareto. Đồ thị Pareto sắp xếp các cột theo thứ tự từ lớn nhất đến nhỏ nhất, giúp dễ đọc và dễ diễn giải hơn.



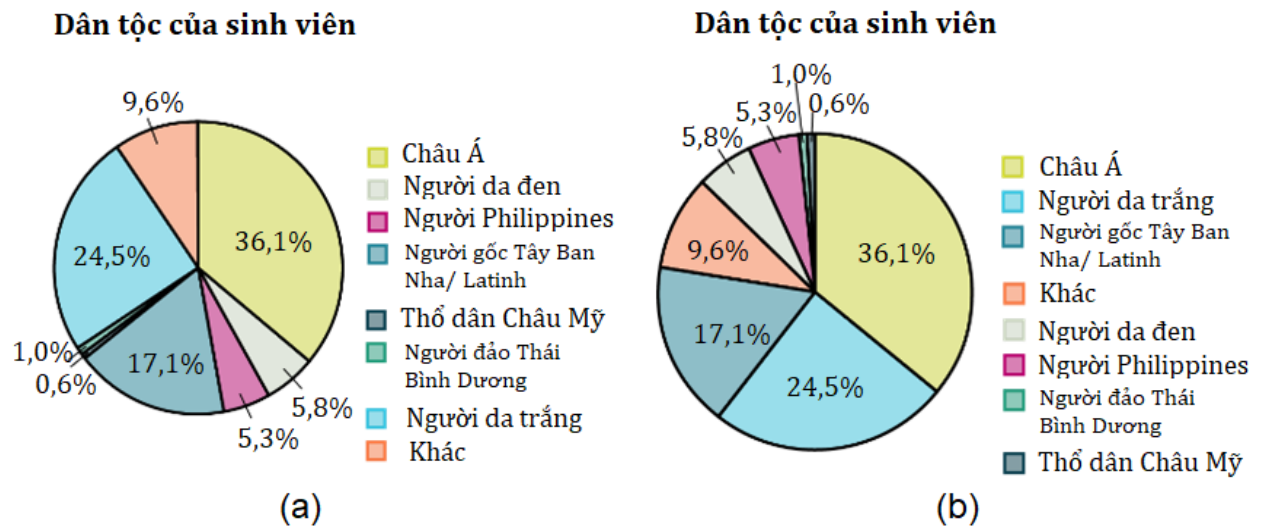
Hình 1.8 Đồ thị thanh với nhóm Khác/Không xác định



Hình 1.9 Đồ thị Pareto với các cột được sắp xếp theo kích thước

Biểu đồ tròn: Không có dữ liệu thiếu

Các biểu đồ tròn dưới đây đã bao gồm nhóm “Khác/Không xác định” (vì tổng các tỷ lệ phần trăm phải cộng lại thành 100%). Biểu đồ trong Hình 1.10(b) được sắp xếp theo kích thước của từng phần quạt, giúp biểu đồ trực quan và dễ hiểu hơn so với biểu đồ không sắp xếp theo thứ tự và sắp xếp theo chữ cái trong Hình 1.10(a).



Hình 1.10

Lấy mẫu²

Việc thu thập thông tin từ toàn bộ quần thể thường tốn kém hoặc gần như không khả thi. Thay vào đó, chúng ta sử dụng một mẫu đại diện cho quần thể. **Mẫu này cần phản ánh đúng các đặc điểm của quần thể mà nó đại diện.** Phần lớn các nhà thống kê áp dụng các phương pháp lấy mẫu ngẫu nhiên để đạt được mục tiêu này. Trong phần này, chúng ta sẽ tìm hiểu một số phương pháp lấy mẫu phổ biến. Có nhiều phương pháp **lấy mẫu ngẫu nhiên (random sampling)** khác nhau, và trong mỗi phương pháp, mỗi thành viên của quần thể ban đầu có cơ hội như nhau để được chọn vào mẫu. Mỗi phương pháp đều có những ưu điểm và hạn chế riêng. Phương pháp cơ bản và dễ hiểu nhất là **lấy mẫu ngẫu nhiên đơn giản (simple random sample)**. Khi áp dụng phương pháp này, bất kỳ nhóm n cá thể nào cũng có khả năng được chọn ngang nhau so với bất kỳ nhóm n cá thể nào khác. Nói cách khác, mọi mẫu có cùng kích thước đều có cơ hội như nhau để được chọn.

Bên cạnh lấy mẫu ngẫu nhiên đơn giản, còn có các phương pháp lấy mẫu khác cũng sử dụng quy trình ngẫu nhiên để chọn mẫu. **Các phương pháp phổ biến bao gồm lấy mẫu phân tầng, lấy mẫu cụm và lấy mẫu hệ thống.**

Để chọn một **mẫu phân tầng (stratified sample)**, đầu tiên, ta chia quần thể thành các nhóm gọi là tầng (strata), sau đó chọn một số phần tử mẫu **tỷ lệ với kích thước** của từng tầng. Ví dụ, bạn có thể chia quần thể sinh viên của trường theo khoa (mỗi khoa là một tầng), rồi chọn mẫu ngẫu nhiên đơn giản theo tỷ lệ từ mỗi khoa để tạo thành mẫu phân tầng. Để thực hiện điều này, bạn đánh số các thành viên trong từng khoa, sau đó sử dụng phương pháp lấy mẫu ngẫu nhiên đơn giản để chọn số phần tử mẫu phù hợp với tỷ lệ từ mỗi khoa. Các thành viên được chọn từ mỗi khoa sẽ tạo thành mẫu phân tầng.

Để chọn một **mẫu cụm (cluster sample)**, đầu tiên, ta chia quần thể thành các cụm (clusters), sau đó ngẫu nhiên chọn một số cụm. Tất cả các cá thể trong các cụm được chọn sẽ tạo thành mẫu cụm. Ví dụ, giả sử bạn muốn lấy mẫu từ quần thể sinh viên của trường và quyết định chọn ngẫu nhiên bốn khoa. Quá trình bắt đầu bằng cách phân chia sinh viên

² Còn gọi là “chọn mẫu”.

toàn trường theo khoa, với mỗi khoa là một cụm. Tiếp theo, bạn sẽ đánh số các khoa và sử dụng phương pháp lấy mẫu ngẫu nhiên đơn giản để chọn ra bốn số khác nhau, mỗi số đại diện cho một khoa. Tất cả sinh viên trong các khoa được chọn sẽ tạo thành mẫu cụm.

Để chọn một **mẫu hệ thống** (systematic sample), ta cần chọn ngẫu nhiên một điểm bắt đầu và sau đó chọn **mỗi phần tử thứ n** từ danh sách quần thể. Ví dụ, giả sử bạn phải thực hiện một cuộc khảo sát qua điện thoại. Danh bạ điện thoại của bạn có chứa 20.000 danh sách nhà ở. Bạn cần chọn 400 tên cho mẫu. Đánh số các phần tử trong quần thể từ 1 đến 20.000, rồi sử dụng phương pháp lấy mẫu ngẫu nhiên đơn giản để chọn một số đại diện cho tên đầu tiên trong mẫu. Sau đó, bạn chọn **mỗi tên thứ 50** cho đến khi có tổng cộng 400 tên (có thể bạn sẽ phải quay lại từ đầu danh sách điện thoại). Phương pháp lấy mẫu hệ thống thường được sử dụng nhờ vào tính đơn giản và dễ dàng áp dụng.

Một loại phương pháp lấy mẫu không ngẫu nhiên là lấy mẫu thuận tiện. **Lấy mẫu thuận tiện** (convenience sampling) liên quan đến việc sử dụng các kết quả có sẵn. Ví dụ, một cửa hàng phần mềm máy tính tiến hành nghiên cứu thị trường bằng cách phỏng vấn những khách hàng tiềm năng đang có mặt tại cửa hàng và đang duyệt qua các phần mềm có sẵn. Kết quả từ phương pháp lấy mẫu thuận tiện có thể rất chính xác trong một số trường hợp, nhưng cũng có thể rất thiên lệch (thiên về một số kết quả nhất định) trong những trường hợp khác.

Việc thu thập dữ liệu cần được thực hiện rất cẩn thận. Thu thập dữ liệu một cách cẩu thả có thể dẫn đến những hậu quả nghiêm trọng. Loại khảo sát bằng thư tín gửi đến các hộ gia đình để họ tự trả lời và gửi lại là loại khảo sát có thể bị thiên lệch rất nhiều (chúng có thể thiên về một nhóm nhất định). Tốt hơn là người thực hiện khảo sát nên chủ động chọn người tham gia mẫu khảo sát.

Lấy mẫu ngẫu nhiên thực sự (true random sampling) cần được thực hiện với **sự thay thế** (with replacement). Nghĩa là, sau khi một thành viên được chọn, thành viên đó sẽ được trả lại vào quần thể và có thể được chọn lại nhiều lần. Tuy nhiên, vì lý do thực tế, trong hầu hết các quần thể, lấy mẫu ngẫu nhiên đơn giản thường được thực hiện **không có sự thay thế** (without replacement). Các cuộc khảo sát thường được thực hiện theo cách này. Điều này có nghĩa là một thành viên của quần thể chỉ có thể được chọn một lần. Hầu hết các mẫu đều được lấy từ các quần thể lớn và mẫu sẽ nhỏ so với quần thể. Vì vậy, **lấy mẫu không có sự thay thế** (sampling without replacement) thực tế tương tự như **lấy mẫu có sự thay thế** (sampling with replacement) vì khả năng chọn cùng một cá thể nhiều lần trong lấy mẫu có sự thay thế là rất thấp.

Nhìn chung, việc lấy mẫu không có sự thay thế thay vì lấy mẫu có sự thay thế chỉ trở thành vấn đề khi quần thể nhỏ.

Khi phân tích dữ liệu, điều quan trọng là phải phân biệt được **sai số lấy mẫu** (sampling errors) và **sai số ngoài lấy mẫu** (nonsampling errors). **Sai số lấy mẫu** có thể xuất hiện do quá trình lấy mẫu thực tế. Ví dụ, mẫu có thể không đủ lớn. **Sai số ngoài lấy mẫu** xuất hiện do các nhân tố không liên quan đến quá trình lấy mẫu gây ra. Ví dụ, một thiết bị đếm bị lỗi có thể gây ra sai số ngoài lấy mẫu.

Trong thực tế, một mẫu sẽ không bao giờ hoàn toàn đại diện cho quần thể, vì vậy luôn có một số sai số lấy mẫu. Theo quy tắc chung, mẫu càng lớn, sai số lấy mẫu càng nhỏ.

Trong thống kê, **thiên lệch lấy mẫu** (sampling bias) xảy ra khi một mẫu được thu thập từ một quần thể mà một số thành viên của quần thể không có khả năng được chọn như các

thành viên khác (nhớ rằng, mỗi thành viên của quần thể phải có cơ hội được chọn như nhau). Khi có thiên lệch lấy mẫu, kết luận về quần thể đang được nghiên cứu có thể bị sai lệch.

Đánh giá phản biện

Chúng ta cần phải đánh giá một cách phản biện các nghiên cứu thống kê mà chúng ta đọc và phân tích chúng trước khi chấp nhận kết quả của các nghiên cứu đó. Một số vấn đề phổ biến cần lưu ý bao gồm:

- **Vấn đề với mẫu:** Một mẫu phải đại diện cho quần thể. Mẫu không đại diện cho quần thể sẽ bị thiên lệch. Mẫu bị thiên lệch, không đại diện cho quần thể, sẽ cho kết quả không chính xác và không hợp lệ.
- **Mẫu tự chọn:** Những phản hồi chỉ được đưa ra bởi những người tự nguyện tham gia, chẳng hạn như các cuộc khảo sát qua điện thoại, thường không đáng tin cậy.
- **Vấn đề về kích thước mẫu:** Các mẫu quá nhỏ có thể không đáng tin cậy. Nếu có thể lấy các mẫu lớn sẽ tốt hơn. Trong một số trường hợp, việc sử dụng mẫu nhỏ là không thể tránh khỏi và vẫn có thể sử dụng để đưa ra kết luận. Ví dụ: thử nghiệm va chạm xe hơi hoặc thử nghiệm y tế cho các bệnh hiếm gặp.
- **Ảnh hưởng quá mức:** Thu thập dữ liệu hoặc đặt câu hỏi theo cách có thể ảnh hưởng đến phản hồi.
- **Không trả lời hoặc từ chối tham gia:** Những phản hồi thu được có thể không còn đại diện cho quần thể. Thường thì những người có quan điểm mạnh mẽ (cả tích cực và tiêu cực) sẽ trả lời khảo sát, điều này có thể làm ảnh hưởng đến kết quả.
- **Quan hệ nhân quả:** Mỗi quan hệ tương quan giữa hai biến không có nghĩa là một biến gây ra sự thay đổi của biến kia. Chúng có thể có liên quan (tương quan) vì mối quan hệ của chúng thông qua một biến khác.
- **Nghiên cứu tự tài trợ hoặc vì lợi ích cá nhân:** Một nghiên cứu được thực hiện bởi một cá nhân hoặc tổ chức nhằm hỗ trợ cho yêu cầu của họ. Liệu nghiên cứu này có khách quan không? Hãy đọc kỹ nghiên cứu để đánh giá công trình. Đừng tự động cho rằng nghiên cứu là tốt, nhưng cũng đừng tự động cho rằng nghiên cứu là tệ. Hãy đánh giá nó dựa trên giá trị và công trình đã thực hiện.
- **Lạm dụng dữ liệu:** Biểu đồ được trình bày không đúng cách, dữ liệu không đầy đủ hoặc thiếu ngữ cảnh.
- **Trộn lẫn:** Khi ảnh hưởng của nhiều yếu tố đối với một phản hồi không thể tách biệt được. Trộn lẫn làm cho việc rút ra kết luận hợp lệ về ảnh hưởng của từng yếu tố trở nên khó khăn hoặc không thể thực hiện được.

Ví dụ 1.11

Vấn đề

Một nghiên cứu được thực hiện để xác định mức học phí trung bình mà sinh viên đại học San Jose State phải trả mỗi học kỳ. Mỗi sinh viên trong các mẫu dưới đây được hỏi về số tiền học phí họ đã trả trong học kỳ Mùa Thu. Hãy xác định loại hình lấy mẫu trong từng trường hợp dưới đây:

- Một mẫu gồm 100 sinh viên đại học San Jose State được chọn bằng cách phân loại danh sách sinh viên theo nhóm năm học (sinh viên năm nhất, năm hai, năm ba, hoặc năm cuối), sau đó lựa chọn 25 sinh viên từ mỗi nhóm.
- Một bộ phát số ngẫu nhiên được sử dụng để chọn một sinh viên từ danh sách theo thứ tự chữ cái của tất cả sinh viên đại học trong học kỳ mùa thu. Bắt đầu từ sinh viên được chọn, cứ mỗi sinh viên thứ 50 sẽ được chọn cho đến khi có tổng cộng 75 sinh viên được đưa vào mẫu.
- Một phương pháp hoàn toàn ngẫu nhiên được sử dụng để chọn ra 75 sinh viên. Mỗi sinh viên đại học trong học kỳ Mùa Thu có cơ hội được chọn như nhau ở bất kỳ giai đoạn nào của quá trình lấy mẫu.
- Các năm học (năm nhất, năm hai, năm ba và năm cuối) được đánh số lần lượt là 1, 2, 3 và 4. Một bộ sinh số ngẫu nhiên được sử dụng để chọn ra hai trong số các năm học đó. Tất cả sinh viên thuộc hai năm học được chọn sẽ trở thành mẫu.
- Một trợ lý hành chính được yêu cầu đứng trước thư viện vào một ngày thứ Tư và hỏi 100 sinh viên đại học đầu tiên mà anh ta gặp về học phí họ đã trả trong học kỳ Mùa Thu. 100 sinh viên này sẽ tạo thành mẫu.

Lời giải

a. Lấy mẫu phân tầng (Stratified); b. Lấy mẫu hệ thống (Systematic); c. Lấy mẫu ngẫu nhiên đơn giản (Simple Random); d. Lấy mẫu cụm (Cluster); e. Lấy mẫu thuận tiện (Convenience)

Bài tập vận dụng 1.11

Bạn sẽ sử dụng bộ phát số ngẫu nhiên để tạo ra các loại mẫu khác nhau từ dữ liệu. Dưới đây là bảng hiển thị sáu bộ điểm bài kiểm tra (mỗi bài kiểm tra có tổng điểm 10) của một lớp học thống kê cơ bản.

Bảng 1.6

#1	#2	#3	#4	#5	#6
5	7	10	9	8	3
10	5	9	8	7	6
9	10	8	6	7	9
9	10	10	9	8	9
7	8	9	5	7	4
9	9	9	10	8	7
7	7	10	9	8	8
8	8	9	10	8	8
9	7	8	7	7	8
8	8	10	9	8	7

Hướng dẫn: Sử dụng Excel để chọn mẫu.

1. Tạo mẫu phân tầng (Stratified Sampling)

Yêu cầu: Chọn ngẫu nhiên 3 điểm bài kiểm tra từ mỗi cột (từ cột #1 đến #6).

Thực hiện:

Bước 1: Đánh số thứ tự từ 1 đến 10 cho mỗi hàng trong bảng (đây là số thứ tự của các điểm bài kiểm tra trong từng cột).

Bước 2: Nhập vào một ô trống của trang tính Excel công thức: `=RANDBETWEEN(1,10)` (Đây là hàm tạo số nguyên ngẫu nhiên từ 1 đến 10)

Bước 3: Ở cột 1, thực hiện các bước sau:

- Nhấn phím F9 để tính lại hàm Excel.
- Ghi lại số trong ô chứa công thức `RANDBETWEEN(1,10)`.
- Lập lại thêm 2 lần để có 3 con số.

Bước 4: Lập lại quy trình bước 3 cho 5 cột còn lại.

Tổng cộng bạn sẽ thu được 3 điểm $\times$ 6 cột = 18 điểm $\rightarrow$ tạo thành mẫu phân tầng.

2. Tạo mẫu cụm (Cluster Sampling)

Yêu cầu: Chọn ngẫu nhiên 2 trong 6 cột.

Thực hiện:

Bước 1: Nhập vào một ô trống của trang tính Excel công thức: `=RANDBETWEEN(1,6)` (Đây là hàm tạo số nguyên ngẫu nhiên từ 1 đến 6)

Bước 2: Nhấn phím F9 để tính lại hàm Excel, ghi l phím F9 để t chứa công thức `RANDBETWEEN(1,6)`

Bước 3: Nhấn phím F9 để tính lại hàm Excel, ghi lại số trong ô chứa công thức `RANDBETWEEN(1,6)`. Nếu số mới trùng với số cũ, thì tiếp tục nhấn F9 để có con số khác số cũ.

Bước 4: Lấy toàn bộ 10 điểm từ mỗi cột tương ứng với con số đã có.

3. Tạo mẫu ngẫu nhiên đơn giản (Simple Random Sampling)

Yêu cầu: Chọn ngẫu nhiên 15 điểm bài kiểm tra từ tổng số 60 điểm.

Thực hiện:

Bước 1: Đánh số thứ tự cho toàn bộ các điểm từ 1 đến 60 (ví dụ: đi theo từng hàng từ trái sang phải). Ví dụ: Ô A1 (cột #1, hàng 1) là số 1, Ô B1 (cột #2, hàng 1) là số 2, ..., Ô F10 (cột #6, hàng 10) là số 60.

Bước 2: Tạo 15 ô trống, mỗi ô nhập công thức: `=RANDBETWEEN(1,60)`

Bước 3: Copy $\rightarrow$ Paste Values để cố định

Bước 4: Kiểm tra các con số có trùng lặp (có thể dùng "Remove Duplicates" trong menu Data. Nếu trùng lặp thì tạo con số mới với `RANDBETWEEN`)

Bước 5: Lấy điểm từ ô tương ứng.

4. Tạo mẫu hệ thống (Systematic Sampling)

Yêu cầu: Chọn 12 điểm trong 60 điểm theo hệ thống cách đều.

Thực hiện:

Bước 1: Tính khoảng cách mẫu: $k = 60 / 12 = 5$

Bước 2: Chọn một số ngẫu nhiên từ 1 đến 5 làm vị trí bắt đầu (r) bằng hàm `RANDBETWEEN(1,5)` trong Excel

Bước 3: Tạo dãy số hệ thống bằng cách từ r, lần lượt cộng thêm 5 cho đến khi đủ 12 số: r, r+5, r+10, ..., r+55

Bước 4: Dùng 12 số thứ tự vừa tạo $\rightarrow$ tra cứu điểm bài kiểm tra tương ứng trong bảng. Danh sách 12 điểm đó là mẫu hệ thống.

Ví dụ 1.12*Vấn đề*

Xác định loại phương pháp lấy mẫu được sử dụng (ngẫu nhiên đơn giản, phân tầng, hệ thống, cụm, hay thuận tiện).

- Một huấn luyện viên bóng đá chọn sáu cầu thủ từ một nhóm các bé trai từ tám đến mười tuổi, bảy cầu thủ từ một nhóm các bé trai từ 11 đến 12 tuổi và ba cầu thủ từ một nhóm các bé trai từ 13 đến 14 tuổi để thành lập một đội bóng đá giải trí.
- Một người thăm dò ý kiến phỏng vấn tất cả nhân viên nhân sự tại năm công ty công nghệ cao khác nhau.
- Một nhà nghiên cứu giáo dục trường trung học phỏng vấn 50 giáo viên trường trung học công lập và 50 giáo viên trường trung học tư thục.
- Một nhà nghiên cứu y tế phỏng vấn mỗi bệnh nhân thứ ba trong danh sách bệnh nhân ung thư tại một bệnh viện địa phương.
- Một cổ vấn trường trung học sử dụng máy tính để tạo ra 50 số ngẫu nhiên và sau đó chọn các học sinh có tên tương ứng với các số đó.
- Một sinh viên phỏng vấn các bạn cùng lớp trong lớp đại số của mình để xác định số lượng trung bình các bộ quần jeans mà một sinh viên sở hữu.

Lời giải

a. Phân tầng; b. Cụm; c. Phân tầng; d. Hệ thống; e. Ngẫu nhiên đơn giản; f. Thuận tiện

Bài tập vận dụng 1.12

Xác định loại phương pháp lấy mẫu được sử dụng (ngẫu nhiên đơn giản, phân tầng, hệ thống, cụm, hay thuận tiện).

Một hiệu trưởng trường trung học thăm dò ý kiến 50 học sinh năm nhất, 50 học sinh năm hai, 50 học sinh năm ba và 50 học sinh năm cuối về những thay đổi chính sách liên quan đến các hoạt động sau giờ học.

Nếu chúng ta kiểm tra hai mẫu đại diện cho cùng một quần thể, ngay cả khi sử dụng các phương pháp lấy mẫu ngẫu nhiên, chúng sẽ không hoàn toàn giống nhau. Giống như có biến thiên trong dữ liệu, cũng có biến thiên trong các mẫu. Khi bạn quen dần với việc lấy mẫu, biến thiên này sẽ bắt đầu trở nên tự nhiên.

LƯU Ý

Trong phần **Khoảng tin cậy** (Confidence Intervals) của giáo trình, các công thức tính kích thước mẫu được cung cấp để xác định số lượng mẫu khi lấy mẫu từ một quần thể. Kích thước mẫu sẽ phụ thuộc vào độ chính xác mong muốn thay vì vào kích thước quần thể. Điều này có thể hơi ngược với suy nghĩ thông thường, vì chúng ta có thể nghĩ rằng kích thước mẫu phải tỷ lệ thuận với kích thước quần thể. Tuy nhiên, thực tế là một mẫu

1.000 người có thể đủ để đại diện cho quần thể 100.000 người cũng như 1.000.000 người, nếu yêu cầu về độ chính xác là giống nhau. Khi áp dụng công thức tính kích thước mẫu trong phần **Khoảng tin cậy**, bạn sẽ nhận thấy rằng kích thước quần thể không phải là yếu tố quyết định trong việc xác định kích thước mẫu.

Ví dụ 1.13

Giả sử trường Cao đẳng ABC có 10.000 sinh viên bán thời gian (quần thể). Chúng ta quan tâm đến số tiền trung bình mà mỗi sinh viên bán thời gian chi tiêu cho sách vở trong kỳ học mùa thu. Việc khảo sát tất cả 10.000 sinh viên là một nhiệm vụ gần như không thể thực hiện được.

Giả sử chúng ta thực hiện hai mẫu khác nhau:

Đầu tiên, chúng ta sử dụng phương pháp lấy mẫu thuận tiện và khảo sát mười sinh viên từ lớp Hóa học hữu cơ kỳ 1. Nhiều sinh viên trong số này đang học môn Giải tích kỳ 1 cùng với lớp Hóa học hữu cơ. Số tiền họ chi tiêu cho sách vở như sau:

\$128; \$87; \$173; \$116; \$130; \$204; \$147; \$189; \$93; \$153

Mẫu thứ hai được lấy từ danh sách những người cao tuổi tham gia lớp thể dục và ta chọn mỗi người cao tuổi thứ năm trong danh sách, tổng cộng là mười người. Số tiền họ chi tiêu cho sách vở là:

\$50; \$40; \$36; \$15; \$50; \$100; \$40; \$53; \$22; \$22

Khả năng có người học xuất hiện trong cả hai mẫu này là rất thấp.

Vấn đề

Bạn có nghĩ rằng một trong hai mẫu này đại diện cho (hoặc đặc trưng cho) toàn bộ quần thể 10.000 sinh viên bán thời gian không?

Lời giải

a. Không. Mẫu đầu tiên có thể chủ yếu bao gồm các sinh viên thiên về khoa học. Ngoài lớp Hóa học hữu cơ, một số sinh viên còn học môn Giải tích kỳ 1. Sách cho các môn này thường rất đắt. Hầu hết các sinh viên này, có khả năng, đang chi tiêu nhiều hơn mức trung bình của sinh viên bán thời gian cho sách vở. Mẫu thứ hai là một nhóm những người cao tuổi, có lẽ đang tham gia các lớp học để nâng cao sức khỏe hoặc thỏa mãn sở thích. Số tiền họ chi cho sách vở có lẽ ít hơn nhiều so với sinh viên bán thời gian trung bình. Cả hai mẫu đều có sự thiên lệch. Thêm vào đó, trong cả hai trường hợp, không phải tất cả sinh viên đều có cơ hội được chọn vào mẫu.

Vấn đề

b. Vì các mẫu này không đại diện cho toàn bộ quần thể, liệu có hợp lý khi sử dụng kết quả của chúng để mô tả toàn bộ quần thể không?

Lời giải

b. Không. Với các mẫu này, mỗi thành viên trong quần thể không có cơ hội chọn lựa như nhau.

Giả sử bây giờ chúng ta lấy một mẫu thứ ba. Chúng ta sẽ chọn 10 sinh viên bán thời gian từ các ngành học như hóa học, toán học, tiếng Anh, tâm lý học, xã hội học, lịch sử, điều

duỡng, giáo dục thể chất, nghệ thuật và phát triển trẻ nhỏ. (Giả định rằng đây là tất cả các ngành học mà sinh viên bán thời gian tại Cao đẳng ABC tham gia và số lượng sinh viên đăng ký trong mỗi ngành học là như nhau.) Mỗi sinh viên được chọn qua phương pháp lấy mẫu ngẫu nhiên đơn giản. Các số ngẫu nhiên được tạo ra bằng máy tính và mỗi sinh viên sẽ được chọn nếu số của họ khớp với số ngẫu nhiên được tạo ra. Các sinh viên trong mẫu chi tiêu các khoản sau:

\$180; \$50; \$150; \$85; \$260; \$75; \$180; \$200; \$200; \$150.

Vấn đề

c. Mẫu này có bị thiên lệch không?

Lời giải

c. Mẫu này không bị thiên lệch, nhưng nếu sử dụng một mẫu lớn hơn, khả năng mẫu này đại diện cho quần thể sẽ cao hơn. Tuy nhiên, đối với các phương pháp lấy mẫu có thiên lệch, dù mẫu có lớn đến đâu, vẫn có nguy cơ mẫu không đại diện cho quần thể.

Sinh viên thường thắc mắc liệu thay vì khảo sát toàn bộ quần thể việc chỉ lấy mẫu có “đủ tốt” không. Nếu quy trình khảo sát được thực hiện đúng cách, thì câu trả lời là có.

Bài tập vận dụng 1.13

Một đài phát thanh địa phương có 20.000 thính giả. Đài muốn biết liệu khán giả của mình sẽ thích nghe nhiều nhạc hơn hay nhiều chương trình trò chuyện hơn. Việc khảo sát tất cả 20.000 thính giả là một nhiệm vụ gần như không thể.

Đài sử dụng phương pháp lấy mẫu thuận tiện và khảo sát 200 người đầu tiên mà họ gặp tại một trong các sự kiện hòa nhạc của đài. 24 người cho biết họ muốn nghe nhiều chương trình trò chuyện hơn, trong khi 176 người cho biết họ muốn nghe nhiều nhạc hơn.

Bạn có nghĩ rằng mẫu này đại diện cho (hoặc đặc trưng cho) toàn bộ quần thể 20.000 thính giả không?

Biến thiên trong dữ liệu

Biến thiên (variation) luôn tồn tại trong bất kỳ tập dữ liệu nào. Chẳng hạn, lon nước ngọt ghi 16 ounce có thể thực tế chứa nhiều hoặc ít hơn con số này. Trong một nghiên cứu, tám lon 16 ounce được đo và cho ra lượng nước (tính bằng ounce) như sau:

15,8; 16,1; 15,2; 14,8; 15,8; 15,9; 16,0; 15,5

Sự sai khác này có thể xuất phát từ việc nhiều người đo lường khác nhau hoặc do nhà sản xuất không rót chính xác 16 ounce chất lỏng vào mỗi lon. Vì vậy, các nhà sản xuất thường xuyên kiểm tra để đảm bảo lượng nước trong lon 16 ounce nằm trong khoảng cho phép.

Khi bạn thu thập dữ liệu, kết quả của bạn có thể hơi khác so với kết quả của người khác dù cùng mục đích—đó là điều hoàn toàn tự nhiên. Tuy nhiên, nếu hai hay nhiều người cùng đo mà thu được kết quả chênh lệch quá lớn, đã đến lúc tất cả cần xem xét lại phương pháp thu thập dữ liệu và độ chính xác của mình.

Biến thiên trong mẫu

Như đã đề cập, hai hoặc nhiều **mẫu** được lấy ngẫu nhiên từ cùng một **quần thể**—dù có đặc điểm gần giống quần thể—vẫn thường khác nhau. Giả sử Doreen và Jung cùng muốn khảo sát thời gian ngủ trung bình mỗi đêm của sinh viên trong trường. Cả hai đều lấy mẫu 500 sinh viên: Doreen dùng lấy mẫu hệ thống, còn Jung dùng lấy mẫu cụm. Mẫu của Doreen chắc chắn sẽ khác mẫu của Jung. Ngay cả khi họ sử dụng cùng một phương pháp, khả năng cao các mẫu vẫn khác nhau—và điều đó hoàn toàn bình thường, không có ai sai.

Hãy suy nghĩ điều gì khiến mẫu của Doreen và Jung khác nhau.

Nếu cả hai tăng kích thước mẫu (nghĩa là lấy nhiều giá trị dữ liệu hơn), kết quả của mẫu (thời gian ngủ trung bình của sinh viên) có thể gần sát hơn với giá trị trung bình thật của quần thể. Tuy vậy, gần như chắc chắn mẫu của họ vẫn khác nhau. Do đó, **biến thiên trong mẫu** là hiện tượng luôn tồn tại.

Kích thước mẫu

Kích thước mẫu—thường ký hiệu là n (số quan sát)—rất quan trọng. Các ví dụ trước đây trong sách đều dùng mẫu nhỏ. Với nhiều mục đích, mẫu chỉ vài trăm quan sát (thậm chí ít hơn) đã đủ. Trong khảo sát dư luận, cỡ mẫu khoảng 1.200–1.500 quan sát được xem là đủ lớn và đáng tin cậy, miễn là quá trình khảo sát ngẫu nhiên và được tiến hành cẩn thận. Sau này, bạn sẽ thấy rằng kích thước nhỏ hơn nữa cũng có thể cho kết quả rất tốt—lý do sẽ rõ khi bạn học về **khoảng tin cậy**.

Cần lưu ý rằng nhiều mẫu lớn vẫn có thể **thiên lệch**. Ví dụ, khảo sát qua các cuộc gọi tự nguyện hầu như luôn bị thiên lệch, vì người tham gia tự quyết định có hồi đáp hay không.

1.3 Thang đo

Khi đã có một tập dữ liệu, bạn cần sắp xếp chúng để phân tích tần suất xuất hiện của từng giá trị. Tuy nhiên, trong quá trình tính tần suất, bạn có thể phải làm tròn kết quả nhằm đạt được độ chính xác nhất có thể.

Các thang đo

Cách thức đo lường một tập dữ liệu được gọi là **thang đo** (level of measurement). Việc áp dụng đúng các phương pháp thống kê phụ thuộc vào việc nhà nghiên cứu nhận diện chính xác mức thang đo; không phải phép toán thống kê nào cũng thích hợp cho mọi loại dữ liệu. Dữ liệu được chia thành bốn thang đo, xếp từ thấp đến cao như sau:

- *Thang đo định danh (Nominal scale level)*
- *Thang đo thứ bậc (Ordinal scale level)*
- *Thang đo khoảng (Interval scale level)*
- *Thang đo tỷ lệ (Ratio scale level)*

Dữ liệu đo trên **thang đo định danh (nominal scale)** là **dữ liệu định tính** (phân loại). Các hạng mục, màu sắc, tên nhãn, món ăn ưa thích cùng với câu trả lời “có” hoặc “không” là ví dụ của dữ liệu mức định danh. Thang đo định danh **không** được sắp xếp theo thứ tự; chẳng hạn, xếp pizza trước sushi không có ý nghĩa.

Các hãng sản xuất điện thoại thông minh cũng là một ví dụ về dữ liệu thang đo định danh. Ở đây, dữ liệu chỉ gồm tên các hãng sản xuất điện thoại thông minh; giữa các thương hiệu này không tồn tại thứ tự chung được thừa nhận, mặc dù mỗi người có thể có sở thích riêng. Dữ liệu thang đo định danh không thể sử dụng trong tính toán.

Dữ liệu được đo bằng **thang đo thứ bậc (ordinal scale)** giống với dữ liệu thang đo định danh, nhưng có một khác biệt quan trọng: dữ liệu thang đo thứ bậc có thể sắp xếp theo thứ tự. Ví dụ về dữ liệu thang đo thứ bậc là danh sách năm công viên quốc gia hàng đầu ở Hoa Kỳ. Năm công viên này có thể được xếp hạng từ một đến năm, nhưng chúng ta không thể đo lường mức chênh lệch giữa các hạng.

Một ví dụ khác về việc sử dụng thang đo thứ bậc là khảo sát về dịch vụ du thuyền, trong đó các câu trả lời cho các câu hỏi về chuyến du lịch là “xuất sắc”, “tốt”, “đạt yêu cầu” và “không đạt yêu cầu”. Những câu trả lời này được xếp theo thứ tự từ mức mong muốn nhất đến mức ít mong muốn nhất. Tuy nhiên, không thể đo được sự khác biệt giữa hai dữ liệu. Giống như dữ liệu thang đo định danh, dữ liệu thang đo thứ bậc không thể sử dụng trong tính toán.

Dữ liệu đo trên **thang đo khoảng (interval scale)** giống dữ liệu thang đo thứ bậc ở chỗ có thứ tự rõ ràng, nhưng khác ở chỗ khoảng cách giữa các giá trị có thể đo được, dù thang đo này không có mốc 0 tuyệt đối.

Những thang đo nhiệt độ như Celsius ($^{\circ}\text{C}$) và Fahrenheit ($^{\circ}\text{F}$) là ví dụ điển hình của thang đo khoảng. Ở cả hai thang đo nhiệt độ này, 40° đúng bằng 100° trừ 60° – chênh lệch này hoàn toàn có ý nghĩa. Tuy nhiên, 0° không phải mức “không còn nhiệt độ”, bởi vẫn tồn tại các giá trị âm như -10°F hay -15°C , và lạnh hơn 0° .

Dữ liệu thang đo khoảng có thể tham gia các phép cộng, trừ, nhưng **không thể so sánh theo tỷ lệ**. Chẳng hạn, 80°C không “nóng gấp bốn lần” 20°C , cũng như 80°F không gấp bốn lần 20°F ; Tỷ lệ 80 trên 20 (hay bốn trên một) không có ý nghĩa.

Dữ liệu đo theo **thang đo tỷ lệ (ratio scale)** khắc phục hoàn toàn vấn đề về tỷ lệ và cung cấp lượng thông tin phong phú nhất. Dữ liệu thang tỷ lệ giống như dữ liệu thang đo khoảng, điểm khác biệt then chốt là có mốc 0 tuyệt đối, nên ta có thể tính toán tỷ lệ giữa các giá trị. Ví dụ, bốn điểm thi trắc nghiệm cuối kỳ môn thống kê là 80, 68, 20 và 92 (thang điểm tối đa 100). Các bài thi được chấm bằng máy. Ta có thể sắp xếp kết quả từ thấp đến cao: 20, 68, 80, 92.

Khoảng cách giữa các điểm có ý nghĩa: 92 cao hơn 68 đúng 24 điểm. Đồng thời, vì có mốc 0 tuyệt đối, phép so sánh tỷ lệ cũng có giá trị: điểm thấp nhất có thể là 0, nên 80 gấp bốn lần 20; nói cách khác, 80 điểm cao gấp bốn lần 20 điểm.

Tần suất

Hai mươi sinh viên được hỏi họ làm việc bao nhiêu giờ mỗi ngày. Câu trả lời của họ (tính theo giờ) như sau: 5; 6; 3; 3; 2; 4; 7; 5; 2; 3; 5; 6; 5; 4; 4; 3; 5; 2; 5; 3.

Bảng 1.7 liệt kê các giá trị dữ liệu theo thứ tự tăng dần cùng tần suất xuất hiện của từng giá trị.

Bảng 1.7 Bảng tần suất giờ làm việc của sinh viên

GIÁ TRỊ DỮ LIỆU	TẦN SUẤT
2	3
3	5
4	3
5	6
6	2
7	1

Tần suất (frequency) là số lần một giá trị xuất hiện trong tập dữ liệu. Theo Bảng 1.7, có ba sinh viên làm việc 2 giờ, năm sinh viên làm việc 3 giờ, v.v. Tổng các giá trị trong cột tần suất bằng 20, đúng bằng tổng số sinh viên trong mẫu.

Tần suất tương đối (relative frequency) là tỷ lệ—dưới dạng phân số, phần trăm hoặc số thập phân—giữa số lần một giá trị xuất hiện so với tổng số quan sát. Để tính tần suất tương đối, chia tần suất của mỗi giá trị cho tổng số sinh viên trong mẫu (ở đây là 20). Ví dụ, giá trị 2 giờ có tần suất 3, nên tần suất tương đối là $3/20$, bằng 0,15 (15 %).

Bảng 1.8 Bảng tần suất giờ làm việc của sinh viên với tần suất tương đối

GIÁ TRỊ DỮ LIỆU	TẦN SUẤT	TẦN SUẤT TƯƠNG ĐỐI
2	3	$\frac{3}{20}$ hoặc 0,15
3	5	$\frac{5}{20}$ hoặc 0,25
4	3	$\frac{3}{20}$ hoặc 0,15
5	6	$\frac{6}{20}$ hoặc 0,30
6	2	$\frac{2}{20}$ hoặc 0,10
7	1	$\frac{1}{20}$ hoặc 0,05

Bảng 1.8 cho thấy tổng các giá trị trong cột tần suất tương đối là $\frac{20}{20}$, hoặc 1.

Tần suất tương đối tích lũy (Cumulative relative frequency) là giá trị cộng dồn của các tần suất tương đối cho đến hàng hiện tại. Để tính tần suất tương đối tích lũy của một hàng, bạn lấy tần suất tương đối của hàng này và cộng với tất cả các tần suất tương đối ở các hàng phía trên, như thể hiện trong Bảng 1.9.

Bảng 1.9 Bảng tần suất giờ làm việc của sinh viên với tần suất tương đối và tích lũy tương đối

GIÁ TRỊ DỮ LIỆU	TẦN SUẤT	TẦN SUẤT TƯƠNG ĐỐI	TẦN SUẤT TƯƠNG ĐỐI TÍCH LŨY
2	3	$\frac{3}{20}$ hoặc 0,15	0,15
3	5	$\frac{5}{20}$ hoặc 0,25	$0,15 + 0,25 = 0,40$
4	3	$\frac{3}{20}$ hoặc 0,15	$0,40 + 0,15 = 0,55$
5	6	$\frac{6}{20}$ hoặc 0,30	$0,55 + 0,30 = 0,85$
6	2	$\frac{2}{20}$ hoặc 0,10	$0,85 + 0,10 = 0,95$
7	1	$\frac{1}{20}$ hoặc 0,05	$0,95 + 0,05 = 1,00$

Mục cuối cùng trong cột tần suất tương đối tích lũy có giá trị bằng 1, cho thấy 100% dữ liệu đã được quy gộp.

LƯU Ý

Do làm tròn, tổng các giá trị trong cột tần suất tương đối không nhất thiết phải đúng bằng 1, và giá trị cuối cùng trong cột tần suất tương đối tích lũy cũng có thể khác 1. Tuy nhiên, cả hai đều nên xấp xỉ 1.

Bảng 1.10 biểu thị chiều cao (tính bằng inch) của một mẫu gồm 100 cầu thủ bóng đá bán chuyên nghiệp.

Bảng 1.10 Bảng tần suất chiều cao của cầu thủ bóng đá

CHIỀU CAO (INCH)	TẦN SUẤT	TẦN SUẤT TƯƠNG ĐỐI	TẦN SUẤT TƯƠNG ĐỐI TÍCH LŨY
59,95–61,95	5	$\frac{5}{100} = 0,05$	0,05
61,95–63,95	3	$\frac{3}{100} = 0,03$	$0,05 + 0,03 = 0,08$
63,95–65,95	15	$\frac{15}{100} = 0,15$	$0,08 + 0,15 = 0,23$
65,95–67,95	40	$\frac{40}{100} = 0,40$	$0,23 + 0,40 = 0,63$
67,95–69,95	17	$\frac{17}{100} = 0,17$	$0,63 + 0,17 = 0,80$
69,95–71,95	12	$\frac{12}{100} = 0,12$	$0,80 + 0,12 = 0,92$
71,95–73,95	7	$\frac{7}{100} = 0,07$	$0,92 + 0,07 = 0,99$
73,95–75,95	1	$\frac{1}{100} = 0,01$	$0,99 + 0,01 = 1,00$
	Tổng cộng = 100	Tổng cộng = 1,00	

Dữ liệu trong bảng này đã được phân nhóm vào các khoảng sau:

- Từ 59,95 đến 61,95 inch
- Từ 61,95 đến 63,95 inch

- Từ 63,95 đến 65,95 inch
- Từ 65,95 đến 67,95 inch
- Từ 67,95 đến 69,95 inch
- Từ 69,95 đến 71,95 inch
- Từ 71,95 đến 73,95 inch
- Từ 73,95 đến 75,95 inch

Trong mẫu này, có năm cầu thủ có chiều cao thuộc khoảng 59,95–61,95 inch, ba cầu thủ thuộc khoảng 61,95–63,95 inch, mười lăm cầu thủ thuộc khoảng 63,95–65,95 inch, bốn mươi cầu thủ thuộc khoảng 65,95–67,95 inch, mười bảy cầu thủ thuộc khoảng 67,95–69,95 inch, mười hai cầu thủ thuộc khoảng 69,95–71,95 inch, bảy cầu thủ thuộc khoảng 71,95–73,95 inch và một cầu thủ thuộc khoảng 73,95–75,95 inch. Tất cả các giá trị chiều cao đều nằm giữa hai mút khoảng, không bằng đúng điểm đầu hay điểm cuối.

Ví dụ 1.14

Vấn đề

Từ Bảng 1.10, hãy tìm tỷ lệ phần trăm số cầu thủ có chiều cao thấp hơn 65,95 inch.

Lời giải

Nếu bạn quan sát hàng thứ nhất, thứ hai và thứ ba, bạn sẽ thấy tất cả các chiều cao trong các khoảng này đều thấp hơn 65,95 inch. Có tổng cộng $5 + 3 + 15 = 23$ cầu thủ có chiều cao thấp hơn 65,95 inch. Do đó, tỷ lệ phần trăm chiều cao thấp hơn 65,95 inch là $\frac{23}{100}$ hoặc 23%. Tỷ lệ phần trăm này chính là giá trị tần suất tương đối tích lũy trong hàng thứ ba.

Bài tập vận dụng 1.14

Bảng 1.11 hiển thị lượng mưa hàng năm tính bằng inch tại một số thị trấn.

Bảng 1.11

Lượng mưa (Inch)	Tần suất	Tần suất tương đối	Tần suất tương đối tích lũy
2,95–4,97	6	$\frac{6}{50} = 0,12$	0,12
4,97–6,99	7	$\frac{7}{50} = 0,14$	$0,12 + 0,14 = 0,26$
6,99–9,01	15	$\frac{15}{50} = 0,30$	$0,26 + 0,30 = 0,56$
9,01–11,03	8	$\frac{8}{50} = 0,16$	$0,56 + 0,16 = 0,72$
11,03–13,05	9	$\frac{9}{50} = 0,18$	$0,72 + 0,18 = 0,90$
13,05–15,07	5	$\frac{5}{50} = 0,10$	$0,90 + 0,10 = 1,00$
	Tổng cộng = 50	Tổng cộng = 1,00	

Từ Bảng 1.11, hãy tìm phần trăm lượng mưa thấp hơn 9,01 inch.

Ví dụ 1.15*Vấn đề*

Từ Bảng 1.10, hãy tìm phần trăm chiều cao nằm trong khoảng từ 61,95 đến 65,95 inch.

Lời giải

Cộng các tần suất tương đối ở hàng thứ hai và thứ ba: $0,03 + 0,15 = 0,18$ hay 18%.

Bài tập vận dụng 1.15

Từ Bảng 1.11, hãy tìm phần trăm lượng mưa nằm trong khoảng từ 6,99 đến 13,05 inch.

Ví dụ 1.16*Vấn đề*

Sử dụng dữ liệu chiều cao của 100 cầu thủ bóng đá bán chuyên nghiệp trong Bảng 1.10. Điền vào chỗ trống và kiểm tra câu trả lời của bạn.

- Tỷ lệ phần trăm chiều cao nằm trong khoảng 67,95 đến 71,95 inch là: ____.
- Tỷ lệ phần trăm chiều cao nằm trong khoảng 67,95 đến 73,95 inch là: ____.
- Tỷ lệ phần trăm chiều cao lớn hơn 65,95 inch là: ____.
- Số lượng cầu thủ trong mẫu có chiều cao từ 61,95 đến 71,95 inch là: ____.
- Chiều cao thuộc loại dữ liệu nào?
- Mô tả cách bạn có thể thu thập dữ liệu chiều cao sao cho phản ánh đặc trưng của toàn bộ cầu thủ bóng đá bán chuyên.

Lưu ý: Bạn đếm tần suất. Để tìm tần suất tương đối, chia tần suất cho tổng số quan sát. Để tìm tần suất tương đối tích lũy, cộng tất cả các tần suất tương đối trước đó với tần suất tương đối của hàng hiện tại.

Lời giải

- 29%
- 36%
- 77%
- 87
- Dữ liệu liên tục định lượng
- Lấy danh sách cầu thủ từ mỗi đội và chọn mẫu ngẫu nhiên đơn giản từ từng đội.

Bài tập vận dụng 1.16

Từ Bảng 1.11, hãy tìm số lượng thị trấn có lượng mưa nằm trong khoảng từ 2,95 đến 9,01 inch.

Bài tập nhóm

Trong lớp học của bạn, hãy cử một bạn tiến hành khảo sát số lượng anh chị em mà mỗi sinh viên trong lớp có. Hãy lập một bảng tần suất từ dữ liệu thu được. Sau đó bổ sung thêm cột tần suất tương đối và cột tần suất tương đối tích lũy. Trả lời các câu hỏi sau:

1. Có bao nhiêu phần trăm sinh viên trong lớp không có anh chị em?
2. Có bao nhiêu phần trăm sinh viên có từ một đến ba anh chị em?
3. Có bao nhiêu phần trăm sinh viên có dưới ba anh chị em?

Ví dụ 1.17

Mười chín người được hỏi về quãng đường (làm tròn đến dặm gần nhất) họ đi làm mỗi ngày. Dữ liệu thu được như sau: 2; 5; 7; 3; 2; 10; 18; 15; 20; 7; 10; 18; 5; 12; 13; 12; 4; 5; 10. Bảng 1.12 được lập dựa trên tập dữ liệu này.

Bảng 1.12 Tần suất quãng đường đi làm

DỮ LIỆU	TẦN SUẤT	TẦN SUẤT TƯƠNG ĐỐI	TẦN SUẤT TƯƠNG ĐỐI TÍCH LŨY
3	3	$\frac{3}{19}$	0,1579
4	1	$\frac{1}{19}$	0,2105
5	3	$\frac{3}{19}$	0,1579
7	2	$\frac{2}{19}$	0,2632
10	3	$\frac{4}{19}$	0,4737
12	2	$\frac{2}{19}$	0,7895
13	1	$\frac{1}{19}$	0,8421
15	1	$\frac{1}{19}$	0,8948
18	1	$\frac{1}{19}$	0,9474
20	1	$\frac{1}{19}$	1,0000

Vấn đề

- a. Bảng này có chính xác không? Nếu không, sai ở đâu?
- b. Đúng hay sai: 3% số người được khảo sát đi làm quãng đường 3 dặm. Nếu sai, câu đúng là gì? Nếu bảng có lỗi, hãy chỉnh sửa.

- c. Có bao nhiêu người (dưới dạng phân số) đi làm quãng đường 5 hoặc 7 dặm?
- d. Có bao nhiêu người (dưới dạng phân số) đi làm:
- từ 12 dặm trở lên?
 - dưới 12 dặm?
 - trong khoảng từ 5 đến 13 dặm (không bao gồm 5 và 13)?

Lời giải

- a. Không. Cột tần suất trong bảng chỉ cộng lại được 18 thay vì 19, nên thiếu dữ liệu. Ngoài ra, một số giá trị trong cột tần suất tương đối tích lũy cũng bị tính sai.
- b. Sai. Tần suất cho quãng đường 3 dặm là 1 người, tương ứng với khoảng 5,26%, không phải 3%. Ngoài ra, dữ liệu cho quãng đường 2 dặm bị bỏ sót, trong khi có 2 người đi quãng đường này. Cột tần suất tương đối tích lũy đúng phải là: 0,1052; 0,1579; 0,2105; 0,3684; 0,4737; 0,6316; 0,7368; 0,7895; 0,8421; 0,9474; 1,0000.
- c. $\frac{5}{19}$
- d. $\frac{7}{19}, \frac{12}{19}, \frac{7}{19}$

Bài tập vận dụng 1.17

Bảng 1.11 thể hiện lượng mưa trung bình hằng năm (tính bằng inch) tại một mẫu các thị trấn. Trong số các thị trấn được khảo sát, có bao nhiêu phần (dưới dạng phân số) nhận được lượng mưa nằm trong khoảng từ 11,03 đến 13,05 inch mỗi năm

Ví dụ 1.18

Bảng 1.13 trình bày dữ liệu về số năm công tác của 70 nhân viên liên bang.

Bảng 1.13

Số năm công tác	Số lượng nhân viên liên bang
24	2
25	1
26	3
27	0
28	4
29	6
30	11
31	12
32	7
33	8
34	6
35	10

Vấn đề

Trả lời các câu hỏi sau:

- Tần suất tích lũy của số năm công tác trong khoảng từ 30 đến 35 năm (bao gồm cả 30 và 35) là bao nhiêu?
- Tần suất tương đối của số năm công tác đúng 30 năm là bao nhiêu?
- Tần suất tương đối của số năm công tác bằng hoặc nhỏ hơn 30 năm là bao nhiêu?
- Tần suất tương đối của số năm công tác bằng hoặc lớn hơn 25 năm là bao nhiêu?

Lời giải

- Tần suất tích lũy là 54.
- 11/70 hoặc 0,157 hoặc 15,7%
- 27/70 hoặc 0,386 hoặc 38,6%
- 68/70 hoặc 0,971 hoặc 97,1%

Bài tập vận dụng 1.18

Bảng 1.14 tổng hợp số vụ tai nạn giao thông đường bộ dẫn đến tử vong tại Hoa Kỳ trong suốt 18 năm.

Bảng 1.14

Năm	Tổng số vụ tai nạn	Năm	Tổng số vụ tai nạn
Năm 1	36.254	Năm 11	38.444
Năm 2	37.241	Năm 12	39.252
Năm 3	37.494	Năm 13	38.648
Năm 4	37.324	Năm 14	37.435
Năm 5	37.107	Năm 15	34.172
Năm 6	37.140	Năm 16	30.862
Năm 7	37.526	Năm 17	30.296
Năm 8	37.862	Năm 18	29.757
Năm 9	38.491	Tổng cộng	653.782
Năm 10	38.477		

Trả lời các câu hỏi sau.

- Tần suất số ca tử vong được ghi nhận từ năm thứ 7 đến năm thứ 11 là bao nhiêu?
- Tỷ lệ phần trăm số ca tử vong xảy ra sau năm thứ 13 là bao nhiêu?
- Tần suất tương đối của số ca tử vong xảy ra trong năm thứ 7 trở về trước là bao nhiêu?
- Tỷ lệ phần trăm số ca tử vong xảy ra trong năm thứ 18 là bao nhiêu?
- Tần suất tương đối tích lũy đến năm thứ 13 là bao nhiêu? Giải thích ý nghĩa của con số này cho bạn biết điều gì về dữ liệu.

1.4 Thiết kế thực nghiệm và đạo đức nghiên cứu

Aspirin có thực sự giúp giảm nguy cơ đau tim không? Liệu một loại phân bón có hiệu quả hơn loại khác trong việc trồng hoa hồng? Một mối có nguy hiểm với người lái xe tương đương với việc lái xe khi say rượu không? Những câu hỏi như vậy thường được giải đáp thông qua các thực nghiệm ngẫu nhiên. Trong phần này, bạn sẽ tìm hiểu những khía cạnh quan trọng trong thiết kế thực nghiệm. Việc thiết kế nghiên cứu đúng cách giúp đảm bảo thu thập được dữ liệu chính xác và đáng tin cậy.

Mục tiêu của một thực nghiệm là nghiên cứu mối quan hệ giữa hai biến. Khi một biến gây ra sự thay đổi ở biến khác, biến đầu tiên được gọi là **biến độc lập (independent variable)** hoặc **biến giải thích (explanatory variable)**, còn biến bị ảnh hưởng gọi là **biến phụ thuộc (dependent variable)** hoặc **biến đáp ứng (response variable)** — tức là mối quan hệ kích thích và phản ứng. Trong thực nghiệm ngẫu nhiên, nhà nghiên cứu sẽ điều chỉnh các giá trị của biến giải thích và đo lường sự thay đổi tương ứng ở biến đáp ứng. Những giá trị khác nhau của biến giải thích này gọi là các **“xử lý” (treatments)**. **Đơn vị thực nghiệm (experimental unit)** là từng đối tượng hoặc cá thể riêng biệt được đo lường.

Giả sử bạn muốn khảo sát hiệu quả của vitamin E trong việc phòng bệnh. Bạn tuyển một nhóm đối tượng và hỏi họ có thường xuyên sử dụng vitamin E hay không. Bạn nhận thấy nhóm dùng vitamin E có sức khỏe trung bình tốt hơn so với nhóm không dùng. Liệu điều này có chứng minh vitamin E thực sự giúp phòng bệnh không? Câu trả lời là không. Bởi có nhiều điểm khác biệt giữa hai nhóm so sánh ngoài việc tiêu thụ vitamin E. Những người dùng vitamin E thường kèm theo nhiều thói quen tốt cho sức khỏe khác như tập thể dục, ăn uống khoa học, bổ sung thêm các loại vitamin khác, và không hút thuốc lá. Bất kỳ yếu tố nào trong số này đều có thể ảnh hưởng đến sức khỏe. Do đó, nghiên cứu như vậy không thể khẳng định vitamin E là yếu tố quyết định trong việc phòng bệnh.

Các biến bổ sung có thể làm nhiễu kết quả nghiên cứu được gọi là **biến ngấm ẩn (lurking variable)**. Để khẳng định rằng biến giải thích chính là nguyên nhân gây ra sự thay đổi ở biến đáp ứng, cần phải tách biệt biến giải thích ra khỏi các nhân tố khác. Nhà nghiên cứu phải thiết kế thực nghiệm sao cho chỉ có một điểm khác biệt duy nhất giữa các nhóm so sánh, đó chính là các xử lý (treatment) đã được định trước. Việc này được thực hiện bằng cách **gán ngẫu nhiên (random assignment)** các đơn vị thực nghiệm vào các nhóm được xử lý (treatment groups). Khi các đối tượng được gán một cách ngẫu nhiên, các biến ngấm ẩn tiềm năng sẽ được phân bố đều giữa các nhóm. Khi đó, điểm khác biệt duy nhất giữa các nhóm chính là yếu tố do nhà nghiên cứu áp đặt. Do vậy, những khác biệt về kết quả quan sát ở biến đáp ứng phải là hệ quả trực tiếp của các xử lý khác nhau. Nhờ cách làm này, thực nghiệm có thể chứng minh được mối quan hệ nhân quả giữa biến giải thích và phản đáp ứng.

Sức mạnh của sự gợi ý có thể tác động đáng kể đến kết quả của một thực nghiệm. Nhiều nghiên cứu cho thấy kỳ vọng của người tham gia có thể quan trọng không kém so với tác dụng thực tế của thuốc. Trong một nghiên cứu về các loại thuốc tăng cường hiệu suất, các nhà nghiên cứu nhận thấy:

“Kết quả cho thấy việc tin rằng mình đã dùng chất đó tạo ra hiệu suất gần tương đương với việc thực sự sử dụng thuốc. Ngược lại, sử dụng thuốc mà không nhận thức được việc sử

dụng thuốc không mang lại sự cải thiện hiệu suất đáng kể.” (McClung, M. Collins, D. “Bởi vì tôi biết nó sẽ như vậy!”: Tác dụng giả dược của một chất hỗ trợ sinh lực đối với hiệu suất thể thao. Tạp chí Tâm lý Thể thao & Thể dục. 29 tháng 6 năm 2007 (3):382-94. Truy cập ngày 30 tháng 4 năm 2013.)

Khi việc tham gia nghiên cứu gây ra phản ứng về mặt thể chất ở người tham gia, rất khó để cô lập ảnh hưởng của biến giải thích. Để khắc phục tác động của sức mạnh gợi ý, các nhà nghiên cứu thường thiết lập một **nhóm đối chứng (control group)**. Nhóm này được nhận một liệu pháp **giả dược (placebo)** — một xử lý không thể ảnh hưởng đến biến đáp ứng. Nhóm đối chứng giúp cân bằng tác động của việc tham gia thực nghiệm với tác động của các **xử lý (treatment)** thực tế. Tất nhiên, nếu bạn tham gia nghiên cứu và biết rõ mình chỉ nhận được viên thuốc không chứa hoạt chất, thì sức mạnh của sự gợi ý không còn ảnh hưởng nữa. Trong các thực nghiệm ngẫu nhiên, kỹ thuật **mù hóa (blinding)** hoặc **mã hóa (masking)** được sử dụng để tránh ảnh hưởng của sự gợi ý. Khi một người tham gia bị mù hóa, họ không biết mình thuộc nhóm nhận xử lý thật hay nhóm nhận giả dược. **Thực nghiệm mù đôi (double-blind experiment)** là trường hợp cả người tham gia lẫn nhà nghiên cứu đều bị mù hóa, không biết ai nhận xử lý thật và ai nhận giả dược.

Ví dụ 1.19

Vấn đề

Quý Nghiên cứu và Điều trị Mùi và Vị đã tiến hành một nghiên cứu nhằm khảo sát xem liệu mùi có ảnh hưởng đến khả năng học tập hay không. Các đối tượng tham gia thực hiện trò chơi giải mê cung nhiều lần trong khi đeo khẩu trang. Họ hoàn thành mê cung trên giấy bút ba lần với khẩu trang có mùi hoa và ba lần với khẩu trang không mùi. Người tham gia được gán ngẫu nhiên để đeo khẩu trang có mùi hoa trong ba lần thử đầu hoặc trong ba lần thử cuối. Ở mỗi lần thử, các nhà nghiên cứu ghi nhận thời gian hoàn thành giải mê cung và đánh giá của người tham gia về mùi khẩu trang: tích cực, tiêu cực hoặc trung lập.

- Hãy mô tả biến giải thích và biến đáp ứng trong nghiên cứu này.
- Các xử lý trong nghiên cứu là gì?
- Xác định các biến ngầm ẩn có thể ảnh hưởng đến kết quả nghiên cứu.
- Có thể áp dụng kỹ thuật mù hóa (blinding) trong nghiên cứu này không?

Lời giải

- Biến giải thích là mùi (scent), biến đáp ứng là thời gian hoàn thành giải mê cung.
- Có hai xử lý: khẩu trang có mùi hoa và khẩu trang không mùi.
- Tất cả các đối tượng đều trải qua cả hai xử lý. Thứ tự các xử lý được gán ngẫu nhiên, do đó không có sự khác biệt giữa các nhóm xử lý. Việc gán ngẫu nhiên này loại trừ ảnh hưởng của các biến ngầm ẩn.
- Người tham gia chắc chắn nhận biết được liệu họ có ngửi thấy mùi hoa hay không, nên không thể áp dụng mù hóa đối với người tham gia. Tuy nhiên, các nhà nghiên cứu

thực hiện đo thời gian có thể được làm mù; vì người quan sát sẽ không biết đối tượng đang sử dụng loại khẩu trang nào.

Bài tập vận dụng 1.19

Nhóm Nghiên cứu Giả dược (Placebo Research Group) đã tiến hành một nghiên cứu nhằm xác định mức độ ảnh hưởng của hiệu ứng giả dược. Một nhóm người được chọn ngẫu nhiên và được yêu cầu làm bài kiểm tra trước và sau khi uống một viên thuốc gây đau đầu nhẹ. Ở một nửa số người được chọn ngẫu nhiên, viên thuốc đó được thay bằng viên thuốc giả (placebo) có hình dạng tương tự nhưng không có tác dụng gì. Ở mỗi lần thử, các nhà nghiên cứu ghi nhận sự thay đổi về thời gian hoàn thành bài kiểm tra trước và sau khi uống thuốc.

- Mô tả biến giải thích và biến đáp ứng trong nghiên cứu này.
 - Các xử lý trong nghiên cứu là gì?
 - Xác định các biến ngấm ẩn có thể ảnh hưởng đến nghiên cứu.
 - Có thể áp dụng kỹ thuật mù hóa (blinding) trong nghiên cứu này không?
-

Thuật ngữ chính

Biến (Variable): Đặc điểm được quan tâm của mỗi người hoặc đối tượng trong quần thể.

Biến đáp ứng (Response Variable): Biến phụ thuộc trong thí nghiệm; giá trị được đo lường sự thay đổi vào cuối thí nghiệm.

Biến giải thích (Explanatory Variable): Biến độc lập trong thí nghiệm; giá trị được các nhà nghiên cứu kiểm soát.

Biến ngầm ẩn (Lurking Variable): Biến có ảnh hưởng đến nghiên cứu mặc dù nó không phải là biến giải thích hay biến đáp ứng.

Biến phân loại (Categorical Variable): Là các biến nhận giá trị dưới dạng tên gọi hoặc nhãn phân loại.

Biến số trị (Numerical Variable): Các biến nhận giá trị được biểu thị bằng con số.

Đơn vị thực nghiệm (Experimental Unit): Bất kỳ cá thể hoặc vật thể nào được đo lường trong nghiên cứu.

Dữ liệu (Data): Một tập hợp các quan sát (tập hợp các kết quả có thể xảy ra); hầu hết dữ liệu có thể chia thành hai nhóm: **định tính** (một thuộc tính mà giá trị được biểu thị bằng nhãn) hoặc **định lượng** (một thuộc tính mà giá trị được biểu thị bằng con số). Dữ liệu định lượng lại được chia thành hai nhóm nhỏ: **rời rạc** và **liên tục**. Dữ liệu rời rạc là kết quả của việc đếm (ví dụ: số học sinh của một dân tộc trong lớp hoặc số lượng sách trên kệ). Dữ liệu liên tục là kết quả của việc đo lường (ví dụ: quãng đường di chuyển hoặc cân nặng hành lý).

Dữ liệu định lượng (Quantitative Data): Một tập hợp các quan sát (tập hợp các kết quả có thể xảy ra); dữ liệu định lượng có thuộc tính được biểu thị bằng con số, được chia thành hai nhóm nhỏ: rời rạc và liên tục. Dữ liệu rời rạc là kết quả của việc đếm (ví dụ: số học sinh thuộc dân tộc nào đó trong lớp hoặc số sách trên kệ). Dữ liệu liên tục là kết quả của việc đo lường (ví dụ: quãng đường di chuyển hoặc trọng lượng hành lý).

Dữ liệu định tính (Qualitative Data): Một tập hợp các quan sát (tập hợp các kết quả có thể xảy ra); dữ liệu định tính có thuộc tính được biểu thị bằng nhãn.

Gán ngẫu nhiên (Random Assignment): Hành động phân chia các đơn vị thực nghiệm vào các nhóm xử lý bằng phương pháp ngẫu nhiên.

Giả dược (Placebo): Một dạng xử lý không có tác động thực sự đến biến giải thích.

Khảo sát (Survey): Một nghiên cứu thu thập dữ liệu do các cá nhân báo cáo.

Lấy mẫu có thay thế (Sampling with Replacement): Khi một thành viên của quần thể được chọn vào mẫu, thành viên đó sẽ được trả lại quần thể trước khi chọn cá thể tiếp theo.

Lấy mẫu hệ thống (Systematic Sampling): Phương pháp chọn mẫu ngẫu nhiên; liệt kê các thành viên của quần thể. Sử dụng lấy mẫu ngẫu nhiên đơn giản để chọn điểm bắt đầu trong danh sách. Gọi $k = (\text{số thành viên trong quần thể}) / (\text{số thành viên cần lấy mẫu})$. Chọn mỗi cá thể thứ k trong danh sách, bắt đầu từ điểm đã chọn ngẫu nhiên. Nếu cần, quay lại đầu danh sách để hoàn thành mẫu.

Lấy mẫu không thay thế (Sampling without Replacement): Một thành viên của quần thể chỉ có thể được chọn vào mẫu một lần duy nhất. Nếu đã được chọn, thành viên đó sẽ không được trả lại vào quần thể trước khi chọn các cá thể tiếp theo.

Lấy mẫu ngẫu nhiên (Random Sampling): Phương pháp chọn mẫu sao cho mọi thành viên trong quần thể đều có cơ hội bằng nhau được chọn.

Lấy mẫu ngẫu nhiên đơn giản (Simple Random Sampling): Phương pháp đơn giản để chọn mẫu ngẫu nhiên; mỗi thành viên trong quần thể được gán một số. Sử dụng bộ phát số ngẫu nhiên để chọn một tập các nhãn số này. Các nhãn được chọn ngẫu nhiên chính là các thành viên thuộc mẫu.

Lấy mẫu phân tầng (Stratified Sampling): Phương pháp chọn mẫu ngẫu nhiên nhằm đảm bảo các nhóm con (phân tầng) trong quần thể được đại diện đầy đủ; chia quần thể thành các nhóm (tầng). Sử dụng lấy mẫu ngẫu nhiên đơn giản để chọn một số cá thể theo tỷ lệ từ mỗi tầng.

Lấy mẫu theo cụm (Cluster Sampling): Phương pháp chọn mẫu ngẫu nhiên bằng cách chia quần thể thành các nhóm (cụm), sau đó sử dụng phương pháp lấy mẫu ngẫu nhiên đơn giản để chọn một số cụm. Tất cả các cá thể trong các cụm được chọn đều được đưa vào mẫu khảo sát.

Lấy mẫu thuận tiện (Convenience Sampling): Phương pháp chọn mẫu không ngẫu nhiên, chọn các cá thể dễ tiếp cận nhất. Phương pháp này có thể dẫn đến dữ liệu thiên lệch.

Mẫu (Sample): Một phần tử con của quần thể được nghiên cứu.

Mẫu đại diện (Representative Sample): Một phần tử con của quần thể có đặc điểm tương tự đặc điểm của quần thể.

Mô hình toán học (Mathematical Models): Mô tả một hiện tượng bằng các khái niệm toán học như phương trình, bất phương trình, phân phối, v.v.

Mù hóa (Blinding): Là phương pháp không cho người tham gia biết họ đang nhận loại xử lý nào trong quá trình nghiên cứu.

Nhóm đối chứng (Control Group): Là nhóm trong một thí nghiệm ngẫu nhiên, không nhận bất kỳ xử lý (treatment) nào, nhưng được quản lý giống hệt các nhóm khác.

Quần thể (Population): Toàn bộ các cá thể, vật thể hoặc phép đo mà các đặc tính của chúng đang được nghiên cứu.

Sai số lấy mẫu (Sampling Error): Sự biến thiên tự nhiên phát sinh khi chọn một mẫu để đại diện cho quần thể lớn hơn; sự biến thiên này giảm khi kích thước mẫu tăng, do đó chọn mẫu lớn hơn sẽ giảm sai số lấy mẫu.

Sai số ngoài lấy mẫu (Nonsampling Error): Vấn đề ảnh hưởng đến độ tin cậy của dữ liệu lấy mẫu nằm ngoài sự biến thiên tự nhiên; bao gồm nhiều lỗi do con người như thiết kế nghiên cứu kém, phương pháp lấy mẫu thiên lệch, thông tin không chính xác do người tham gia cung cấp, lỗi nhập liệu và phân tích kém.

Tần suất (Frequency): Số lần một giá trị dữ liệu xuất hiện.

Tần suất tương đối (Relative Frequency): Tỷ số giữa số lần một giá trị xuất hiện trong tập hợp kết quả và tổng số kết quả trong toàn bộ không gian mẫu.

Tần suất tương đối tích lũy (Cumulative Relative Frequency): Áp dụng cho một tập hợp các quan sát được sắp xếp theo thứ tự từ nhỏ đến lớn. Tần suất tương đối tích lũy là tổng tần suất tương đối của tất cả các giá trị nhỏ hơn hoặc bằng giá trị đã cho.

Tham số (Parameter): Một con số dùng để đại diện cho đặc điểm của quần thể, thường không dễ xác định trực tiếp.

Thiên lệch lấy mẫu (Sampling Bias): Trường hợp không phải tất cả thành viên trong quần thể đều có cơ hội được chọn như nhau.

Thông số thống kê (Statistic): Đặc trưng số của mẫu; một thống kê dùng để ước lượng tham số tương ứng của quần thể.

Thực nghiệm mù đôi (Double-blind experiment): Một thí nghiệm trong đó cả đối tượng tham gia lẫn các nhà nghiên cứu làm việc với đối tượng đều không biết thông tin về xử lý (treatment) được áp dụng.

Trung bình (Average): Còn được gọi là mean (số bình quân) hoặc arithmetic mean (bình quân số học); là một con số mô tả tính hướng tâm của dữ liệu.

Tỷ lệ (Proportion): Là tỷ lệ, tức phần trăm số trường hợp thuộc nhóm quan tâm trên tổng số thuộc mẫu.

Xác suất (Probability): Một số nằm trong khoảng từ 0 đến 1, bao gồm cả 0 và 1, thể hiện khả năng xảy ra của một sự kiện cụ thể.

Xử lý (Treatments): Các giá trị hoặc thành phần khác nhau của biến giải thích được áp dụng trong thí nghiệm.

Tóm tắt chương

1.1 Định nghĩa về thống kê, xác suất và thuật ngữ chính

Lý thuyết toán học về thống kê sẽ dễ học hơn khi bạn hiểu được ngôn ngữ chuyên ngành. Phần này trình bày các thuật ngữ quan trọng sẽ được sử dụng xuyên suốt trong toàn bộ tài liệu.

1.2 Dữ liệu, Lấy mẫu và Biến thiên trong Dữ liệu và Lấy mẫu

Dữ liệu là các mục thông tin riêng lẻ được lấy từ quần thể hoặc mẫu. Dữ liệu có thể được phân loại thành dữ liệu định tính (phân loại), dữ liệu định lượng liên tục hoặc dữ liệu định lượng rời rạc.

Do không thực tế khi đo lường toàn bộ quần thể trong một nghiên cứu, các nhà nghiên cứu sử dụng mẫu để đại diện cho quần thể. Mẫu ngẫu nhiên là nhóm đại diện được chọn từ quần thể bằng phương pháp đảm bảo mỗi cá thể trong quần thể đều có cơ hội bằng nhau để được đưa vào mẫu. Các phương pháp lấy mẫu ngẫu nhiên bao gồm lấy mẫu ngẫu nhiên đơn giản, lấy mẫu phân tầng, lấy mẫu theo cụm và lấy mẫu hệ thống. Lấy mẫu thuận tiện là phương pháp không ngẫu nhiên, thường cho kết quả dữ liệu có thiên lệch.

Các mẫu khác nhau chứa các cá thể khác nhau sẽ cho ra dữ liệu khác nhau, ngay cả khi các mẫu được chọn kỹ và đại diện cho quần thể. Khi được lựa chọn đúng cách, mẫu lớn hơn sẽ mô phỏng quần thể chính xác hơn mẫu nhỏ. Có nhiều vấn đề tiềm ẩn có thể ảnh hưởng đến độ tin cậy của mẫu. Dữ liệu thống kê cần được phân tích một cách phản biện, không chỉ đơn giản là chấp nhận.

1.3 Thang đo

Một số phép tính có thể cho ra kết quả chính xác “giả tạo”. Nếu dữ liệu ban đầu chỉ đo được đến một chữ số thập phân, thì không cần thiết phải tính toán kết quả chính xác đến tám chữ số thập phân. Thông thường, kết quả cuối cùng nên được làm tròn sao cho có nhiều hơn một chữ số thập phân so với dữ liệu gốc. Ví dụ, nếu dữ liệu đo chính xác đến một chữ số thập phân, thì kết quả thống kê cuối cùng nên làm tròn đến hai chữ số thập phân.

Bên cạnh việc làm tròn kết quả, bạn có thể đo lường dữ liệu của mình theo bốn loại thang đo sau đây:

- **Thang đo danh nghĩa (Nominal scale):** Dữ liệu không thể được sắp xếp thứ tự và không thể dùng để tính toán.
- **Thang đo thứ bậc (Ordinal scale):** Dữ liệu có thể sắp xếp thứ tự nhưng không thể đo được sự khác biệt giữa các giá trị.
- **Thang đo khoảng (Interval scale):** Dữ liệu có thứ tự rõ ràng nhưng không có điểm gốc; khoảng cách có thể đo được nhưng không thể tính tỷ số.
- **Thang đo tỷ lệ (Ratio scale):** Dữ liệu có điểm gốc, có thể sắp xếp thứ tự; khoảng cách và tỷ số đều có ý nghĩa.

Khi tổ chức dữ liệu, việc biết một giá trị xuất hiện bao nhiêu lần là rất quan trọng. Có bao nhiêu sinh viên ngành thống kê học từ năm giờ trở lên để chuẩn bị cho kỳ thi? Bao nhiêu phần trăm gia đình trong khu phố của chúng ta nuôi hai con thú cưng? Tần suất, tần suất tương đối và tần suất tương đối tích lũy là các phép đo dùng để trả lời những câu hỏi như vậy.

1.4 Thiết kế thực nghiệm và đạo đức nghiên cứu

Một nghiên cứu được thiết kế kém sẽ không tạo ra dữ liệu đáng tin cậy. Một số thành phần then chốt luôn có trong mọi thực nghiệm. Để loại bỏ biến ngầm ẩn, các đơn vị thực nghiệm phải được gán ngẫu nhiên vào các nhóm xử lý khác nhau. Một trong các nhóm xử lý đó phải là nhóm đối chứng, nhằm thể hiện điều gì xảy ra khi không áp dụng xử lý thực sự. Người tham gia trong nhóm đối chứng nhận một giả dược có hình thức giống hệt các xử lý thực nhưng không ảnh hưởng đến biến đáp ứng. Để đảm bảo tính trung thực của giả dược, cả nhà nghiên cứu và các đối tượng tham gia có thể được mù hóa. Khi một nghiên cứu được thiết kế đúng cách, điểm khác biệt duy nhất giữa các nhóm xử lý chính là biến giải thích do nhà nghiên cứu đặt ra. Do đó, khi các nhóm phản ứng khác nhau với các xử lý khác nhau, sự khác biệt đó chắc chắn do ảnh hưởng của biến giải thích.

“Vấn đề đạo đức phát sinh khi bạn cân nhắc một hành động mang lại lợi ích cho bạn hoặc một mục đích bạn ủng hộ, nhưng lại gây tổn hại hoặc làm giảm lợi ích của người khác, đồng thời vi phạm một số quy tắc.” (Andrew Gelman, “Open Data and Open Methods,” Ethics and Statistics, <http://www.stat.columbia.edu/~gelman/research/published/>)

ChanceEthics1.pdf (truy cập ngày 1 tháng 5 năm 2013).) Vi phạm đạo đức trong thống kê không phải lúc nào cũng dễ dàng nhận ra. Các hiệp hội chuyên môn và cơ quan liên bang thường công bố các hướng dẫn về hành vi ứng xử đúng đắn. Việc bạn nắm vững các quy trình thống kê cơ bản là rất quan trọng để có thể nhận biết được phân tích dữ liệu đúng chuẩn.

Tài liệu tham khảo

1.1 Định nghĩa thống kê, xác suất và các thuật ngữ chính

The Data and Story Library, <http://lib.stat.cmu.edu/DASL/Stories/CrashTestDummies.html> (accessed May 1, 2013).

1.2 Dữ liệu, lấy mẫu và biến thiên trong dữ liệu và lấy mẫu

Gallup-Healthways Well-Being Index. <http://www.well-beingindex.com/default.asp> (accessed May 1, 2013).

2 THỐNG KÊ MÔ TẢ



Hình 2.1 Khi bạn có một lượng lớn dữ liệu, bạn sẽ cần tổ chức chúng theo một cách hợp lý. Trong Hình là những lá phiếu bầu cũ được cuộn lại theo sự giống nhau của chúng để giữ chúng ngăn nắp. (Nguồn: William Greenson/ Wikimedia Commons)

Giới thiệu

Sau khi bạn đã thu thập dữ liệu, bạn sẽ làm gì với chúng? Dữ liệu có thể được mô tả và trình bày bằng nhiều định dạng khác nhau. Ví dụ, giả sử bạn quan tâm đến việc mua một căn nhà trong một khu vực cụ thể. Bạn có thể không có manh mối gì về giá nhà, vì vậy bạn có thể yêu cầu đại lý bất động sản cung cấp cho bạn một bộ dữ liệu mẫu về giá cả. Việc xem xét tất cả các mức giá trong mẫu thường rất khó khăn. Một cách tốt hơn có thể là xem xét giá trung vị và sự biến thiên của giá cả. Trung vị và sự biến thiên chỉ là hai cách mà bạn sẽ được học để mô tả dữ liệu. Đại lý của bạn cũng có thể cung cấp cho bạn một biểu đồ của dữ liệu.

Trong chương này, bạn sẽ nghiên cứu các cách sử dụng công cụ số học và đồ thị để mô tả và hiển thị dữ liệu của mình. Lĩnh vực này của thống kê học được gọi là **"Thống kê mô tả"**. Bạn sẽ học cách tính toán, và quan trọng hơn nữa, cách diễn giải các phép đo và biểu đồ này.

Biểu đồ thống kê là một công cụ giúp bạn tìm hiểu về hình dạng hoặc phân phối của một mẫu hoặc đám đông. So với tập lớn các con số, biểu đồ là một cách hiệu quả hơn để trình bày dữ liệu vì chúng ta có thể thấy ở đâu các dữ liệu tập trung và ở đâu chỉ có một vài giá trị dữ liệu. Báo chí và Internet sử dụng biểu đồ để hiển thị xu hướng và cho phép độc giả

so sánh các sự kiện và con số một cách nhanh chóng. Các nhà thống kê thường vẽ biểu đồ dữ liệu trước tiên để có cái nhìn tổng quan về dữ liệu. Sau đó, các công cụ hình thức hơn có thể được áp dụng.

Một số loại biểu đồ được sử dụng để tóm tắt và tổ chức dữ liệu là đồ thị chấm, đồ thị thanh, đồ thị histogram, đồ thị thân và lá, đa giác tần suất (một loại đồ thị đường gãy khúc), biểu đồ tròn và đồ thị hộp. Trong chương này, chúng ta chú trọng về đồ thị histogram và đồ thị hộp bên cạnh tìm hiểu sơ lược về đồ thị thân và lá, đồ thị đường, đồ thị thanh, đa giác tần suất, và đồ thị chuỗi thời gian. Trọng tâm của chương sẽ là đồ thị histogram và đồ thị hộp.

2.1 Trình bày dữ liệu

Đồ thị thân-lá, đồ thị đường và đồ thị thanh

Đồ thị thân-lá (stem-and-leaf graph hay stemplot), một loại đồ thị đơn giản có nguồn gốc từ lĩnh vực phân tích dữ liệu khám phá (exploratory data analysis). Sử dụng đồ thị thân và lá là lựa chọn tốt cho các tập dữ liệu nhỏ. Để tạo đồ thị, chia mỗi quan sát dữ liệu thành **thân** (stem) và **lá** (leaf). Lá bao gồm **chữ số có nghĩa cuối cùng**. Ví dụ, điểm dữ liệu 23 có thân là 2 và lá là 3. Số 432 có thân là 43 và lá là 2. Tương tự, số 5.432 có thân là 543 và lá là 2. Số thập phân 9,3 có thân là 9 và lá là 3. Để vẽ đồ thị thân và lá, đầu tiên viết các số thân theo hàng dọc từ nhỏ nhất đến lớn nhất. Tiếp theo vẽ một đường thẳng đứng ở bên phải các thân. Sau đó viết các số lá theo thứ tự tăng dần bên cạnh thân tương ứng của chúng.

Ví dụ 2.1

Sau đây là điểm của bài kiểm tra số 1 của lớp Toán tiền giải tích của giáo sư Dean (xếp theo thứ tự tăng dần):

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

Bảng 2.1 Đồ thị thân-lá

Thân	Lá
3	3
4	2 9 9
5	3 5 5
6	1 3 7 8 8 9 9
7	2 3 4 8
8	0 3 8 8 8
9	0 2 4 4 4 4 6
10	0

Đồ thị thân và lá cho thấy hầu hết điểm số nằm trong khoảng 60, 70, 80 và 90. 8 trong số 31 điểm số, hay khoảng $\left(\frac{8}{31}\right) = 26\%$ nằm trong khoảng 90 hoặc 100, một số lượng điểm A khá cao.

Bài tập vận dụng 2.1

Sau đây là điểm số của 30 trận đấu gần nhất của đội bóng rổ Park City (xếp từ theo thứ tự tăng dần):

32; 32; 33; 34; 38; 40; 42; 42; 43; 44; 46; 47; 47; 48; 48; 48; 49; 50; 50; 51; 52; 52; 52; 53; 54; 56; 57; 57; 60; 61

Xây dựng đồ thị thân và lá cho dữ liệu.

Đồ thị thân - lá là một cách nhanh chóng để vẽ đồ thị dữ liệu và cho một hình ảnh chính xác về dữ liệu. Khi xem xét đồ thị thân và lá, bạn nên tìm kiếm một mẫu hình tổng thể của dữ liệu và các **điểm ngoại vi** (outlier). Điểm ngoại vi là một quan sát dữ liệu không phù hợp với phần còn lại của dữ liệu. Đôi khi nó được gọi là **trị cực đoan** (extreme value). Khi bạn vẽ đồ thị, điểm ngoại vi sẽ là điểm xuất hiện không phù hợp với mẫu hình của đồ thị. Một số điểm ngoại vi là kết quả của sai sót (ví dụ, viết 50 thay vì 500) trong khi những điểm khác là chỉ dấu có điều gì đó bất thường đang xảy ra. Để giải thích các điểm ngoại vi, chúng ta cần thêm thông tin, vì vậy chúng ta sẽ đề cập đến chúng chi tiết hơn ở phần sau.

Ví dụ 2.2

Dữ liệu là khoảng cách (tính bằng kilômét) từ nhà đến các siêu thị địa phương. Tạo đồ thị thân và lá cho dữ liệu:

1,1; 1,5; 2,3; 2,5; 2,7; 3,2; 3,3; 3,3; 3,5; 3,8; 4,0; 4,2; 4,5; 4,5; 4,7; 4,8; 5,5; 5,6; 6,5; 6,7; 12,3

Vấn đề

Dữ liệu có vẻ như có sự tập trung nào của các giá trị không?

Lời giải

Giá trị 12.3 có thể là một giá trị ngoại vi. Các giá trị dường như tập trung ở ba và bốn kilômét.

Bảng 2.2

Thân	Lá
1	1 5
2	3 5 7
3	2 3 3 5 8
4	0 2 5 5 7 8
5	5 6
6	5 7
7	
8	
9	
10	
11	
12	3

Lưu ý

Các lá ở bên phải dấu thập phân.

Bài tập vận dụng 2.2

Dữ liệu sau đây cho thấy khoảng cách (tính bằng dặm) từ nhà của các sinh viên lớp thống kê (ngoại trừ) đến trường. Tạo đồ thị thân và lá sử dụng dữ liệu và xác định bất kỳ điểm ngoại vi nào:

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

Ví dụ 2.3

Vấn đề

Đồ thị thân và lá song song cho phép so sánh hai tập dữ liệu trong hai cột. Trong đồ thị thân và lá song song, hai tập lá chia sẻ cùng một thân. Các lá ở bên trái và bên phải

của các thân. Bảng 2.3 và Bảng 2.4 cho thấy tuổi của các tổng thống Mỹ khi nhậm chức và khi qua đời. Xây dựng biểu đồ thân-lá song song cho dữ liệu này.

Bảng 2.3 Tuổi các tổng thống Mỹ khi Nhậm chức

Tổng thống	Tuổi	Tổng thống	Tuổi	Tổng thống	Tuổi
Washington	57	Lincoln	52	Hoover	54
J. Adams	61	A. Johnson	56	F. Roosevelt	51
Jefferson	57	Grant	46	Truman	60
Madison	57	Hayes	54	Eisenhower	62
Monroe	58	Garfield	49	Kennedy	43
J. Q. Adams	57	Arthur	51	L. Johnson	55
Jackson	61	Cleveland	47	Nixon	56
Van Buren	54	B. Harrison	55	Ford	61
W. H. Harrison	68	Cleveland	55	Carter	52
Tyler	51	McKinley	54	Reagan	69
Polk	49	T. Roosevelt	42	G.H.W. Bush	64
Taylor	64	Taft	51	Clinton	47
Fillmore	50	Wilson	56	G. W. Bush	54
Pierce	48	Harding	55	Obama	47
Buchanan	65	Coolidge	51		

Bảng 2.4 Tuổi tổng thống Mỹ khi qua đời

Tổng thống	Tuổi	Tổng thống	Tuổi	Tổng thống	Tuổi
Washington	67	Lincoln	56	Hoover	90
J. Adams	90	A. Johnson	66	F. Roosevelt	63
Jefferson	83	Grant	63	Truman	88
Madison	85	Hayes	70	Eisenhower	78
Monroe	73	Garfield	49	Kennedy	46
J. Q. Adams	80	Arthur	56	L. Johnson	64
Jackson	78	Cleveland	71	Nixon	81
Van Buren	79	B. Harrison	67	Ford	93
W. H. Harrison	68	Cleveland	71	Reagan	93
Tyler	71	McKinley	58		
Polk	53	T. Roosevelt	60		
Taylor	65	Taft	72		
Fillmore	74	Wilson	67		
Pierce	64	Harding	57		
Buchanan	77	Coolidge	60		

*Lời giải***Bảng 2.5**

Tuổi khi Nhậm chức		Tuổi khi Qua đời
9 9 8 7 7 7 6 3 2	4	6 9
8 7 7 7 7 6 6 6 5 5 5 5 4 4 4 4 2 2 1 1 1 1 0	5	3 6 6 7 7 8
9 8 5 4 4 2 1 1 1 0	6	0 0 3 3 4 4 5 6 7 7 7 8
	7	0 1 1 1 2 3 4 7 8 8 9
	8	0 1 3 5 8
	9	0 0 3 3

Bài tập vận dụng 2.3

Bảng cho thấy số trận thắng và thua của đội Atlanta Hawks trong 42 mùa giải. Tạo đồ thị thân và lá song song cho dữ liệu này.

Bảng 2.6

Thua	Thắng	Mùa giải	Thua	Thắng	Mùa giải
34	48	1	41	41	22
34	48	2	39	43	23
46	36	3	44	38	24
46	36	4	39	43	25
36	46	5	25	57	26
47	35	6	40	42	27
51	31	7	36	46	28
53	29	8	26	56	29
51	31	9	32	50	30
41	41	10	19	31	31
36	46	11	54	28	32
32	50	12	57	25	33
51	31	13	49	33	34
40	42	14	47	35	35
39	43	15	54	28	36
42	40	16	69	13	37
48	34	17	56	26	38
32	50	18	52	30	39
25	57	19	45	37	40
32	50	20	35	47	41
30	52	21	29	53	42

Một loại đồ thị khác hữu ích cho các giá trị dữ liệu cụ thể là **đồ thị đường** (line graph). Ví dụ 2.4 là một đồ thị đường, **trục x** (trục hoành) chứa **trị dữ liệu** và **trục y** (trục tung) chứa **điểm tần suất**. Các điểm tần suất được kết nối bằng các đoạn thẳng.

Bài tập vận dụng 2.3

Bảng sau cho thấy số trận thắng và thua của đội Atlanta Hawks trong 42 mùa giải. Tạo đồ thị thân và lá song song cho dữ liệu này.

Bảng 2.6

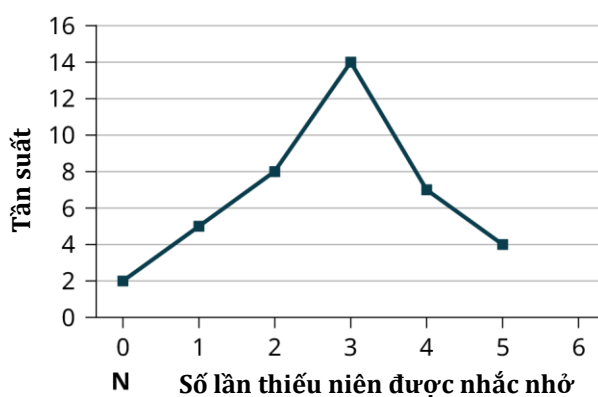
Thua	Thắng	Mùa giải	Thua	Thắng	Mùa giải
34	48	1	41	41	22
34	48	2	39	43	23
46	36	3	44	38	24
46	36	4	39	43	25
36	46	5	25	57	26
47	35	6	40	42	27
51	31	7	36	46	28
53	29	8	26	56	29
51	31	9	32	50	30
41	41	10	19	31	31
36	46	11	54	28	32
32	50	12	57	25	33
51	31	13	49	33	34
40	42	14	47	35	35
39	43	15	54	28	36
42	40	16	69	13	37
48	34	17	56	26	38
32	50	18	52	30	39
25	57	19	45	37	40
32	50	20	35	47	41
30	52	21	29	53	42

Ví dụ 2.4

Trong một cuộc khảo sát, 40 phụ huynh được hỏi một thiếu niên được nhắc nhở làm việc nhà bao nhiêu lần mỗi tuần. Kết quả được hiển thị trong Bảng 2.7 và trong Hình 2.2.

Bảng 2.7

Số lần thiếu niên được nhắc nhở	Tần suất
0	2
1	5
2	8
3	14
4	7
5	4



Hình 2.2

Bài tập vận dụng 2.4

Trong một cuộc khảo sát, 40 người được hỏi bao nhiêu lần mỗi năm họ đưa xe vào cửa hàng sửa chữa. Kết quả được hiển thị trong Bảng 2.8. Xây dựng đồ thị đường.

Bảng 2.8

Số lần vào cửa hàng	Tần suất
0	7
1	10
2	14
3	9

Đồ thị thanh (Bar graphs) bao gồm các thanh được tách biệt với nhau. Các thanh có thể là hình chữ nhật hoặc có thể là hộp chữ nhật (được sử dụng trong đồ thị ba chiều), và

chúng có thể nằm dọc hoặc ngang. **Đồ thị thanh** được hiển thị trong Ví dụ 2.5 có nhóm tuổi được biểu diễn trên **trục x** và tỷ lệ trên **trục y**.

Ví dụ 2.5

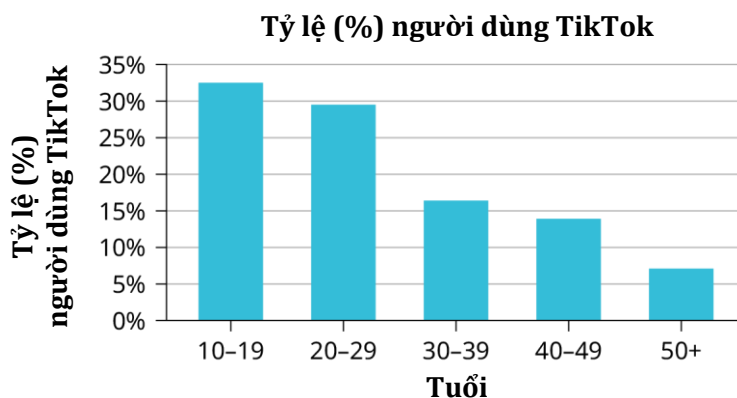
Vấn đề

Tỷ lệ phần trăm theo tuổi của người dùng TikTok sống tại Mỹ được hiển thị trong Bảng 2.9. Xây dựng đồ thị thanh cho dữ liệu này.

Bảng 2.9

Nhóm tuổi	Tỷ lệ (%) người dùng TikTok
10-19	32,5%
20-29	29,5%
30-39	16,4%
40-49	13,9%
50+	7,1%

Lời giải



Hình 2.3

Bài tập vận dụng 2.5

Dân số ở Park City được thống kê theo nhóm trẻ em, người lớn trong độ tuổi lao động và người nghỉ hưu. Bảng 2.10 cho thấy ba nhóm tuổi, số người trong thị trấn từ mỗi nhóm tuổi và tỷ lệ (%) người trong mỗi nhóm tuổi. Xây dựng đồ thị thanh hiển thị các tỷ lệ.

Bảng 2.10

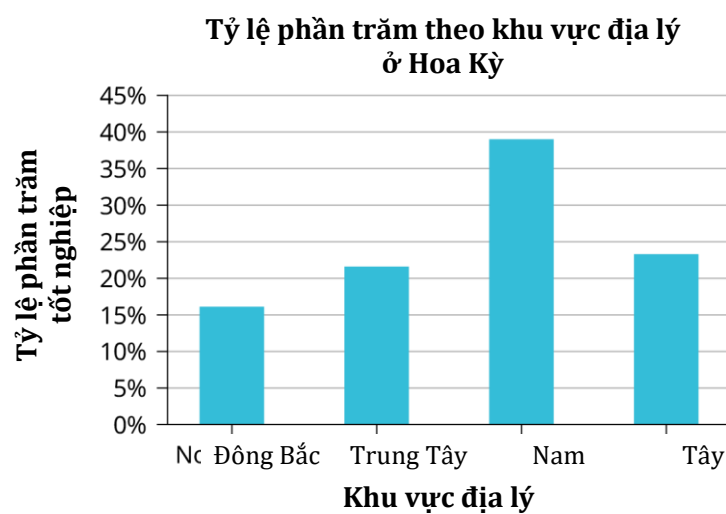
Nhóm tuổi	Số người	Tỷ lệ dân số
Trẻ em	67.059	19%
Người lớn tuổi lao động	152.198	43%
Người nghỉ hưu	131.662	38%

Ví dụ 2.6**Vấn đề**

Các cột trong Bảng 2.11 cho thấy dữ liệu dự báo cho năm 2030 về số lượng và tỷ lệ phần trăm người dân tốt nghiệp trung học theo khu vực địa lý ở Hoa Kỳ. Tạo đồ thị thanh cho dữ liệu này theo khu vực địa lý (dữ liệu định tính) trên trục x và tỷ lệ phần trăm dữ liệu tốt nghiệp trung học (dữ liệu định lượng) trên trục y.

Bảng 2.11

Khu vực	Số người tốt nghiệp	Tỷ lệ phần trăm người tốt nghiệp
Đông Bắc	517.720	16,1%
Trung Tây	695.170	21,6%
Nam	1.253.540	39,0%
Tây	749.400	23,3%

Lời giải**Hình 2.4**

Bài tập vận dụng 2.6

Park city được chia thành sáu khu vực bầu cử. Bảng sau cho thấy tỷ lệ phần trăm của tổng số cử tri đăng ký sống ở mỗi khu vực, cũng như tỷ lệ phần trăm của tổng số dân sống ở mỗi khu vực. Xây dựng đồ thị thanh hiển thị số cử tri đăng ký theo khu vực.

Bảng 2.12

Khu vực	Số cử tri được đăng ký	Tổng số dân thành phố
1	15,5%	19,4%
2	12,2%	15,6%
3	9,8%	9,0%
4	17,4%	18,5%
5	22,8%	20,7%
6	22,3%	16,8%

Ví dụ 2.7

Vấn đề

Dưới đây là bảng hai chiều hiển thị các loại thú cưng được nuôi bởi nam và nữ:

Bảng 2.13

	Chó	Mèo	Cá	Tổng
Nam	4	2	2	8
Nữ	4	6	2	12
Tổng	8	8	4	20

Dựa trên dữ liệu này, hãy tính phân phối có điều kiện cho nhóm nam giới nuôi từng loại thú cưng.

Lời giải

$$\text{Nam giới nuôi chó} = \frac{4}{8} = 0,5$$

$$\text{Nam giới nuôi mèo} = \frac{2}{8} = 0,25$$

$$\text{Nam giới sở hữu cá} = \frac{2}{8} = 0,25$$

Lưu ý: Tổng của tất cả các phân phối có điều kiện phải bằng một. Trong trường hợp này, $0,5 + 0,25 + 0,25 = 1$; do đó, lời giải "đúng".

Bài tập vận dụng 2.7

Dựa trên dữ liệu trong Bảng 2.13, hãy tính phân phối có điều kiện cho nhóm phụ nữ nuôi từng loại thú cưng.

Đồ thị histogram, đa giác tần suất và đồ thị chuỗi thời gian

Đối với hầu hết công việc bạn thực hiện trong cuốn sách này, bạn sẽ sử dụng đồ thị histogram để hiển thị dữ liệu. Một ưu điểm của đồ thị histogram là nó có thể dễ dàng hiển thị các tập dữ liệu lớn. Có một quy tắc là ta nên sử dụng đồ thị histogram khi tập dữ liệu có 100 điểm trị trở lên.

Đồ thị histogram (histogram) bao gồm các hộp xếp liền kề nhau (tiếp giáp). Nó có cả trục hoành và trục tung. Trên trục hoành có các nhãn tên cho những gì dữ liệu đại diện (ví dụ, khoảng cách từ nhà đến trường). Trục tung được ghi nhãn với **tần suất** (frequency) hoặc **tần suất tương đối** (relative frequency), hoặc **phần trăm tần suất** (percent frequency) hoặc **xác suất** (probability). Đồ thị sẽ có cùng hình dạng với cả hai loại nhãn (tần suất hay tần suất tương đối). Đồ thị histogram (giống như stemplot) có thể cung cấp cho ta hình dạng của dữ liệu, tâm và độ phân tán của dữ liệu.

Tần suất tương đối bằng tần suất của một giá trị quan sát của dữ liệu chia cho tổng số các giá trị dữ liệu trong mẫu. (Nhớ rằng: tần suất được định nghĩa là số lần một câu trả lời xuất hiện.) Nếu:

- f = tần suất
- n = tổng số các giá trị dữ liệu (hoặc tổng của các tần suất riêng lẻ), và
- RF = tần suất tương đối,

thì:

$$RF = \frac{f}{n}$$

Ví dụ, giả sử lớp tiếng Anh có 40 học sinh, trong đó 3 học sinh có điểm từ 90% đến 100%, thì $f = 3$, $n = 40$, và $RF = \frac{f}{n} = \frac{3}{40} = 0,075$. 7,5% học sinh nhận được 90-100%. 90-100% là các số đo định lượng.

Để xây dựng đồ thị histogram, trước tiên ta cần quyết định trên đồ thị sẽ có bao nhiêu **thanh** (bar) hoặc **khoảng** (interval), còn được gọi là **lớp** (class), đại diện cho dữ liệu. Nhiều đồ thị histogram có từ 5 đến 15 thanh/lớp để rõ ràng. Số lượng thanh cần được chọn. Hãy chọn điểm bắt đầu cho khoảng đầu tiên nhỏ hơn giá trị dữ liệu nhỏ nhất. **Điểm bắt đầu thuận tiện** (convenient starting point) nên là một giá trị nhỏ hơn trị dữ liệu nhỏ nhất, với số chữ số thập phân nhiều hơn so với trị dữ liệu có nhiều chữ số thập phân nhất. Ví dụ, nếu trị dữ liệu có nhiều chữ số thập phân nhất là 6,1 và đây là giá trị nhỏ nhất, điểm bắt đầu thuận tiện là 6,05 ($6,1 - 0,05 = 6,05$). Chúng ta nói rằng 6,05 có độ chính xác cao hơn. Nếu trị dữ liệu có nhiều chữ số thập phân nhất là 2,23 và trị dữ liệu nhỏ nhất là 1,5, điểm bắt đầu thuận tiện là 1,495 ($1,5 - 0,005 = 1,495$). Nếu trị dữ liệu có nhiều chữ số thập phân nhất là 3,234 và trị dữ liệu nhỏ nhất là 1,0, điểm bắt đầu thuận tiện là 0,9995 ($1,0 - 0,0005 = 0,9995$). Nếu tất cả dữ liệu tình cờ là số nguyên và giá trị nhỏ nhất là 2, thì điểm bắt đầu thuận tiện là 1,5 ($2 - 0,5 = 1,5$). Ngoài ra, với việc điểm bắt đầu và các điểm biên khác của khoảng có nhiều hơn một chữ số thập phân, sẽ không có giá trị dữ liệu nào sẽ rơi

trúng điểm biên. Hai ví dụ tiếp theo đi vào chi tiết về cách xây dựng đồ thị histogram sử dụng dữ liệu liên tục và cách tạo đồ thị histogram sử dụng dữ liệu rời rạc.

Ví dụ 2.8

Dữ liệu sau đây là chiều cao (tính bằng inch đến nửa inch gần nhất) của 100 cầu thủ bóng đá bán chuyên nghiệp. Chiều cao là dữ liệu **liên tục**, vì chiều cao đo lường được.

60; 60,5; 61; 61; 61,5;

63,5; 63,5; 63,5;

64; 64; 64; 64; 64; 64; 64; 64,5; 64,5; 64,5; 64,5; 64,5; 64,5; 64,5; 64,5;

66; 66; 66; 66; 66; 66; 66; 66; 66; 66; 66,5; 66,5; 66,5; 66,5; 66,5; 66,5; 66,5; 66,5;

66,5; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67; 67,5; 67,5; 67,5; 67,5; 67,5; 67,5;

68; 68; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69; 69,5; 69,5; 69,5; 69,5;

70; 70; 70; 70; 70; 70; 70,5; 70,5; 70,5; 71; 71; 71;

72; 72; 72; 72,5; 72,5; 73; 73,5;

74

Giá trị dữ liệu nhỏ nhất là 60. Vì dữ liệu có nhiều chữ số thập phân nhất có một chữ số thập phân (ví dụ, 61,5), chúng ta muốn điểm bắt đầu của mình có hai chữ số thập phân. Vì các số 0,5; 0,05; 0,005; v.v. là những số thuận tiện, ta sử dụng 0,05 và trừ nó từ 60, giá trị nhỏ nhất, cho điểm bắt đầu thuận tiện.

$60 - 0,05 = 59,95$ có độ chính xác cao hơn, chẳng hạn như 61,5 bằng một chữ số thập phân. Điểm bắt đầu là 59,95.

Giá trị lớn nhất là 74, vậy $74 + 0,05 = 74,05$ là giá trị kết thúc.

Tiếp theo, tính chiều rộng của mỗi thanh hoặc khoảng lớp. Để tính chiều rộng này, trừ điểm bắt đầu từ giá trị kết thúc và chia cho số thanh (bạn phải chọn số thanh bạn mong muốn). Giả sử bạn chọn tám thanh.

$$74,05 - 59,95 = 14,1$$

$$14,1 \div 8 = 1,76 \rightarrow 2$$

Lưu ý: Ở đây Chúng ta sẽ làm tròn lên thành 2 và làm cho mỗi thanh (khoảng) rộng 2 đơn vị. Làm tròn lên thành 2 là một cách để ngăn chặn một giá trị rơi trúng biên của thanh. Làm tròn đến con số tiếp theo thường là cần thiết cho dù nó trái các quy tắc làm tròn tiêu chuẩn. Đối với ví dụ này, việc sử dụng 1,76 làm chiều rộng cũng có thể thực hiện. Một hướng dẫn ta có thể sử dụng để tính số lượng thanh cho đồ thị histogram là lấy căn bậc hai của số điểm dữ liệu và sau đó làm tròn đến số nguyên gần nhất (nếu cần thiết). Ví dụ, nếu ta có 150 giá trị dữ liệu, lấy căn bậc hai của 150 (được 12,24) và làm tròn thành 12; ta cần 12 thanh.

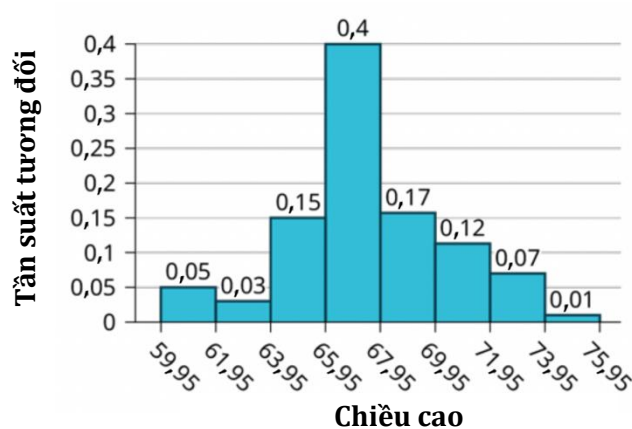
Các điểm biên của khoảng:

- 59,95
- $59,95 + 2 = 61,95$
- $61,95 + 2 = 63,95$

- $63,95 + 2 = 65,95$
- $65,95 + 2 = 67,95$
- $67,95 + 2 = 69,95$
- $69,95 + 2 = 71,95$
- $71,95 + 2 = 73,95$
- $73,95 + 2 = 75,95$

Chiều cao từ 60 đến 61,5 inch nằm trong khoảng 59,95-61,95. Chiều cao 63,5 nằm trong khoảng 61,95-63,95. Chiều cao từ 64 đến 64,5 nằm trong khoảng 63,95-65,95. Chiều cao từ 66 đến 67,5 nằm trong khoảng 65,95-67,95. Chiều cao từ 68 đến 69,5 nằm trong khoảng 67,95-69,95. Chiều cao từ 70 đến 71 nằm trong khoảng 69,95-71,95. Chiều cao từ 72 đến 73,5 nằm trong khoảng 71,95-73,95. Chiều cao 74 nằm trong khoảng 73,95-75,95.

Đồ thị histogram sau đây hiển thị chiều cao trên trục x và tần suất tương đối trên trục y.



Hình 2.5

Bài tập vận dụng 2.8

Dữ liệu sau đây là kích thước giày của 50 học sinh. Kích thước giày là dữ liệu liên tục vì kích thước được đo. Xây dựng đồ thị histogram và tính chiều rộng của mỗi thanh (hoặc khoảng, lớp). Giả sử bạn chọn sáu thanh.

9; 9; 9,5; 9,5; 10; 10; 10; 10; 10; 10; 10; 10,5; 10,5; 10,5; 10,5; 10,5; 10,5; 10,5;
 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11; 11,5; 11,5; 11,5; 11,5; 11,5; 11,5;
 12; 12; 12; 12; 12; 12; 12; 12,5; 12,5; 12,5; 12,5; 14

Ví dụ 2.9

Tạo đồ thị histogram cho dữ liệu sau: số sách được mua bởi 50 sinh viên bán thời gian tại Đại học ABC. Số sách là **dữ liệu rời rạc** (discrete data), vì số lượng sách được đếm.

1; 1; 1; 1; 1; 1; 1; 1; 1; 1;
 2; 2; 2; 2; 2; 2; 2; 2; 2;
 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3; 3;

4; 4; 4; 4; 4; 4;

5; 5; 5; 5; 5;

6; 6

Mười một học sinh mua 1 cuốn sách. Mười học sinh mua 2 cuốn sách. Mười sáu học sinh mua 3 cuốn sách. Sáu học sinh mua 4 cuốn sách. Năm học sinh mua 5 cuốn sách. Hai học sinh mua 6 cuốn sách.

Vì dữ liệu là số nguyên, ta trừ 0,5 từ số 1 (giá trị dữ liệu nhỏ nhất) và cộng 0,5 vào số 6 (giá trị dữ liệu lớn nhất). Khi đó điểm đầu là 0,5 và giá trị cuối là 6,5.

Vấn đề

Tiếp theo, tính chiều rộng của mỗi thanh (khoảng, lớp). Nếu dữ liệu là rời rạc và không có quá nhiều các trị khác nhau, thuận tiện nhất là ta chọn chiều rộng của thanh sao cho các trị dữ liệu nằm giữa thanh. Vì dữ liệu bao gồm các số 1, 2, 3, 4, 5, 6, và điểm đầu là 0,5, chiều rộng của tkhoảng bằng 1 sẽ đặt số 1 ở giữa khoảng từ 0,5 đến 1,5, số 2 ở giữa khoảng từ 1,5 đến 2,5, số 3 ở giữa khoảng từ 2,5 đến 3,5, số 4 ở giữa khoảng từ _____ đến _____, số 5 ở giữa khoảng từ _____ đến _____, và số _____ ở giữa khoảng từ _____ đến _____.

Lời giải

- 3,5 đến 4,5
- 4,5 đến 5,5
- 6
- 5,5 đến 6,5

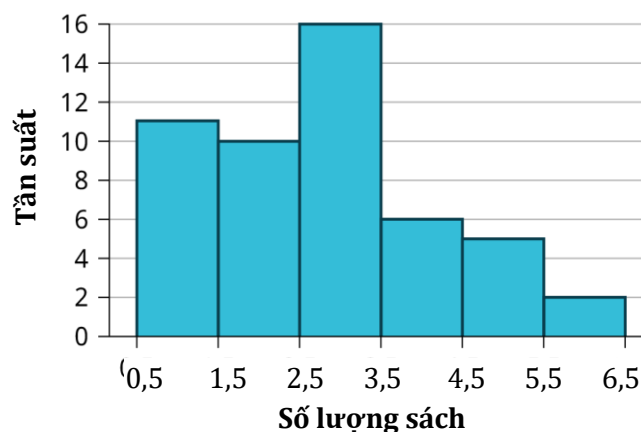
Tính số thanh như sau:

$$6,5 - 0,5 = 6$$

$$6 \div 1 = 6$$

trong đó 1 là chiều rộng của một thanh. Do đó, thanh = 6.

Đồ thị histogram sau đây hiển thị số sách trên trục x và tần suất trên trục y.



Hình 2.6

Bài tập vận dụng 2.9

Dữ liệu sau đây là số môn thể thao được chơi bởi 50 sinh viên. Số môn thể thao là dữ liệu rời rạc vì các môn thể thao được đếm.

1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1; 1;

2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2; 2;

3; 3; 3; 3; 3; 3; 3; 3; 3;

Có 20 sinh viên chơi 1 môn thể thao; 22 sinh viên chơi 2 môn thể thao; 8 sinh viên chơi 3 môn thể thao.

Điền vào chỗ trống cho câu sau. Vì dữ liệu bao gồm các số 1, 2, 3, và điểm bắt đầu là 0,5, chiều rộng khoảng bằng 1 sẽ đặt số 1 ở giữa khoảng 0,5 đến ____, số 2 ở giữa khoảng từ ____ đến ____, và số 3 ở giữa khoảng từ ____ đến ____.

Ví dụ 2.10

Vấn đề

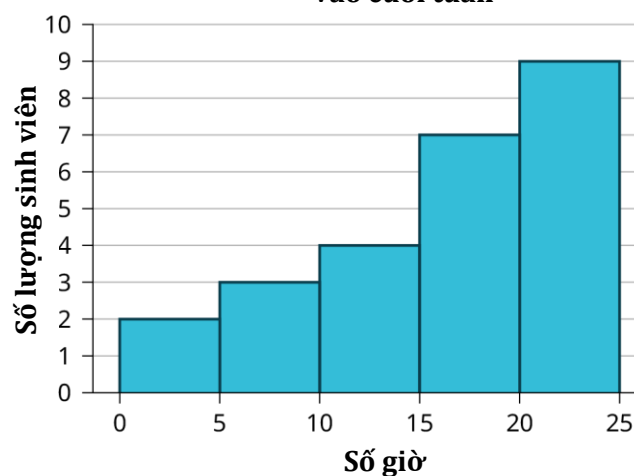
Sử dụng tập dữ liệu này, xây dựng đồ thị histogram.

Bảng 2.14

Số giờ các bạn cùng lớp tôi dành để chơi trò chơi điện tử vào cuối tuần				
9,95	10	2,25	16,75	0
19,5	22,5	7,5	15	12,75
5,5	11	10	20,75	17,5
23	21,9	24	23,75	18
20	15	22,9	18,8	20,5

Lời giải

Số giờ dành cho việc chơi điện tử vào cuối tuần



Hình 2.7

Một số giá trị trong tập dữ liệu này rơi vào biên của khoảng. Thông thường một giá trị sẽ được tính cho khoảng nếu nó rơi vào biên trái của khoảng, và không được tính nếu nó rơi vào biên phải. Tuy vậy các nhà nghiên cứu khác nhau có thể có quy ước đồ thị histogram cho cùng một dữ liệu theo những cách khác nhau. Có nhiều hơn một cách đúng để thiết lập đồ thị histogram³.

Bài tập vận dụng 2.10

Dữ liệu sau đây đại diện cho số nhân viên tại các nhà hàng ở Thành phố New York. Sử dụng dữ liệu này để tạo đồ thị histogram.

22; 35; 15; 26; 40; 28; 18; 20; 25; 34; 39; 42; 24; 22; 19; 27; 22; 34; 40; 20; 38; và 28

Sử dụng 10-19 làm khoảng đầu tiên.

Bài tập nhóm

Đếm số tiền trong túi hoặc ví của bạn; ghi lại các số tiền. Cùng nhau xây dựng đồ thị histogram để hiển thị dữ liệu. Hãy thảo luận về số lượng khoảng mà bạn nghĩ là phù hợp. Bạn có thể muốn thử nghiệm với các số lượng khoảng khác nhau.

LƯU Ý

Sử dụng Excel và R vẽ đồ thị histogram

Excel: sử dụng Histogram trong bộ công cụ Data Analysis.

R/ggplot2 package: sử dụng hàm `ggplot()+ geom_histogram()`

Đa giác tần suất

Đa giác tần suất tương tự như đồ thị đường, và cũng như đồ thị đường, đa giác tần suất giúp cho dữ liệu liên tục dễ hiểu qua trực quan.

Để xây dựng đa giác tần suất, trước tiên hãy kiểm tra dữ liệu và quyết định số khoảng (lớp) được sử dụng trên trục x và trục y. Sau khi chọn các khoảng phù hợp, bắt đầu vẽ các điểm dữ liệu. Sau khi tất cả các điểm được vẽ, vẽ các đoạn thẳng để kết nối chúng.

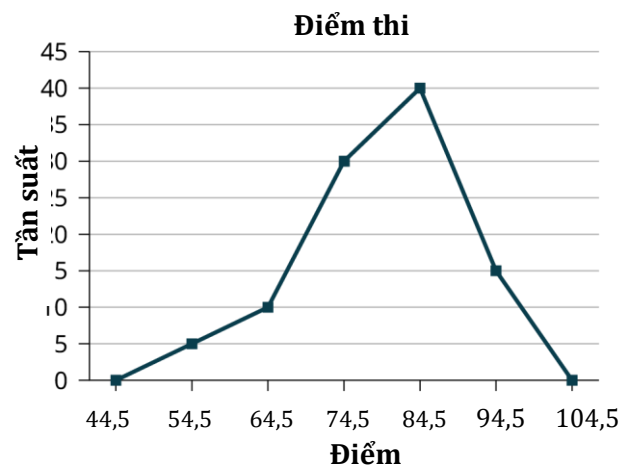
Ví dụ 2.11

Một đa giác tần suất được xây dựng cho bảng tần suất bên dưới.

³ Ta cần lưu ý các phần mềm cũng có quy ước khác nhau để đếm trị dữ liệu theo khoảng.

Bảng 2.15

Phân phối tần suất cho điểm thi cuối kỳ môn giải tích			
Biên dưới	Biên trên	Tần suất	Tần suất tích lũy
49,5	59,5	5	5
59,5	69,5	10	15
69,5	79,5	30	45
79,5	89,5	40	85
89,5	99,5	15	100



Hình 2.8

Nhãn đầu tiên trên trục x là 44,5. Nó đại diện cho một khoảng rộng từ 39,5 đến 49,5. Vì điểm kiểm tra nhỏ nhất là 54,5, khoảng này được dùng chỉ để đồ thị chạm vào trục x. Điểm có nhãn 54,5 đại diện cho khoảng tiếp theo, hoặc khoảng "thực" đầu tiên từ bảng, và chứa năm trị dữ liệu. Các khoảng còn lại tuân theo cùng lập luận với điểm 104,5 đại diện cho khoảng từ 99,5 đến 109,5. Một lần nữa, khoảng này không chứa dữ liệu và được dùng chỉ để đồ thị chạm vào trục x. Nhìn vào đồ thị, chúng ta nói rằng phân phối này bị lệch vì một bên của đồ thị không đối xứng với bên kia.

Bài tập vận dụng 2.11

Xây dựng đa giác tần suất cho tuổi các Tổng thống Hoa Kỳ khi nhậm chức được hiển thị trong Bảng 2.16.

Bảng 2.16

Tuổi khi Nhậm chức	Tần suất
41,5-46,5	4
46,5-51,5	11
51,5-56,5	14
56,5-61,5	9
61,5-66,5	4
66,5-71,5	2

Đa giác tần suất hữu ích để so sánh các phân phối. Việc so sánh được thực hiện bằng cách phủ chồng các đa giác tần suất của các tập dữ liệu khác nhau.

Ví dụ 2.12

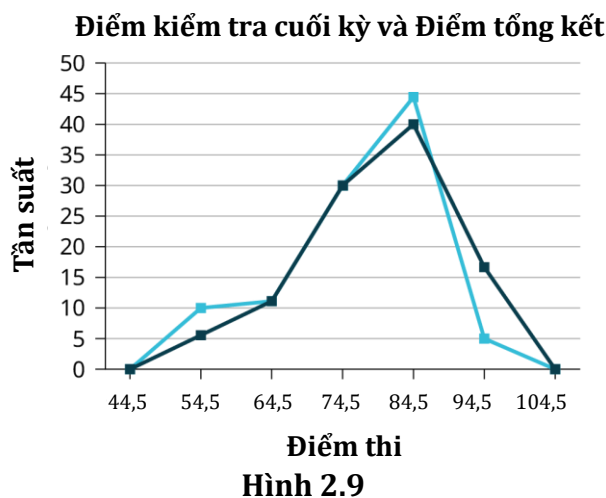
Chúng ta sẽ xây dựng đa giác tần suất phủ để so sánh điểm số từ Ví dụ 2.11 với điểm hết môn giải tích.

Bảng 2.17

Phân phối tần suất cho điểm thi cuối kỳ môn giải tích			
Biên dưới	Biên trên	Tần suất	Tần suất tích lũy
49,5	59,5	5	5
59,5	69,5	10	15
69,5	79,5	30	45
79,5	89,5	40	85
89,5	99,5	15	100

Bảng 2.18

Phân phối tần suất cho điểm hết môn giải tích			
Biên dưới	Biên trên	Tần suất	Tần suất tích lũy
49,5	59,5	10	10
59,5	69,5	10	20
69,5	79,5	30	50
79,5	89,5	45	95
89,5	99,5	5	100



Bài tập vận dụng 2.12

Chúng ta sẽ xây dựng đa giác tần suất phủ lên so sánh điểm số từ Ví dụ 2.12 với điểm thi cuối kỳ môn đại số.

Bảng 2.19

Phân phối tần suất cho điểm thi môn đại số			
Biên dưới	Biên trên	Tần suất	Tần suất tích lũy
49,5	59,5	10	10
59,5	69,5	5	15
69,5	79,5	40	55
79,5	89,5	35	90
89,5	99,5	10	100

Xây dựng đồ thị chuỗi thời gian

Giả sử chúng ta muốn nghiên cứu về khoảng nhiệt độ trong tháng của một vùng. Mỗi ngày vào buổi trưa chúng ta ghi lại nhiệt độ và viết điều này xuống nhật ký. Nhiều nghiên cứu thống kê có thể được thực hiện với dữ liệu này. Chúng ta có thể tìm số bình quân hoặc số trung vị nhiệt độ trong tháng. Chúng ta có thể xây dựng đồ thị histogram hiển thị số ngày mà nhiệt độ nằm trong một khoảng giá trị nhất định. Tuy nhiên, tất cả các phương pháp này bỏ qua một phần dữ liệu mà chúng ta đã thu thập.

Một đặc điểm của dữ liệu mà chúng ta có thể muốn xem xét là thời gian. Vì mỗi ngày được ghép nối với số đo nhiệt độ cho ngày đó, chúng ta không nên nghĩ về dữ liệu như là ngẫu nhiên. Thay vào đó, chúng ta có thể sử dụng thời gian đã cho để áp đặt thứ tự thời gian trên dữ liệu. Đồ thị nhận dạng thứ tự này và hiển thị nhiệt độ thay đổi theo diễn tiến của tháng được gọi là đồ thị chuỗi thời gian.

Để xây dựng đồ thị chuỗi thời gian, chúng ta phải nhìn vào cả hai phần của tập dữ liệu bất cặp của mình. Chúng ta bắt đầu với hệ tọa độ Descartes tiêu chuẩn. Trục hoành được sử dụng để vẽ ngày tháng hoặc giá số thời gian, và trục tung được sử dụng để vẽ các giá trị của biến mà chúng ta đang đo. Bằng cách này, chúng ta gán mỗi điểm trên đồ thị tương ứng với một ngày và một đại lượng đo. Các điểm trên đồ thị thường được kết nối bằng đường thẳng theo thứ tự chúng xuất hiện.

Ví dụ 2.13

Vấn đề

Dữ liệu sau đây cho thấy Chỉ số giá tiêu dùng hàng năm, mỗi tháng, trong mười năm. Xây dựng đồ thị chuỗi thời gian chỉ cho dữ liệu Chỉ số giá tiêu dùng hàng năm.

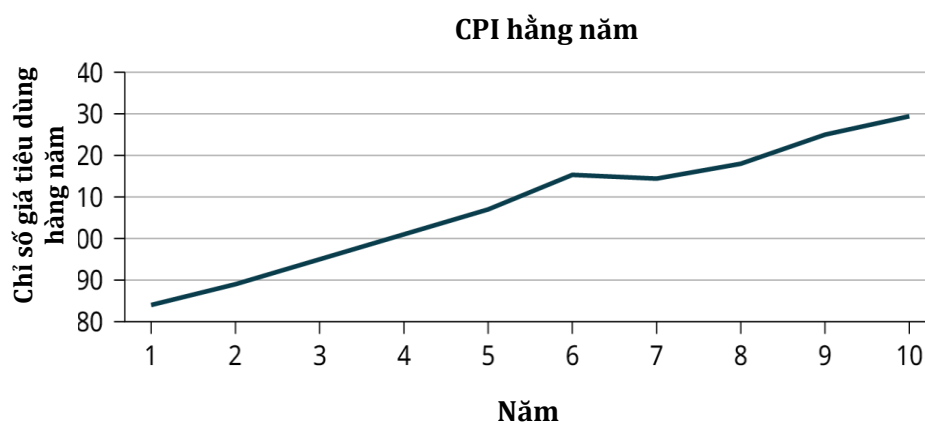
Bảng 2.20

Năm	Tháng 1	Tháng 2	Tháng 3	Tháng 4	Tháng 5	Tháng 6	Tháng 7
1	181,7	183,1	184,2	183,8	183,5	183,7	183,9
2	185,2	186,2	187,4	188,0	189,1	189,7	189,4
3	190,7	191,8	193,3	194,6	194,4	194,5	195,4
4	198,3	198,7	199,8	201,5	202,5	202,9	203,5
5	202,416	203,499	205,352	206,686	207,949	208,352	208,299
6	211,080	211,693	213,528	214,823	216,632	218,815	219,964
7	211,143	212,193	212,709	213,240	213,856	215,693	215,351
8	216,687	216,741	217,631	218,009	218,178	217,965	218,011
9	220,223	221,309	223,467	224,906	225,964	225,722	225,922
10	226,665	227,663	229,392	230,085	229,815	229,478	229,104

Bảng 2.21

Năm	Tháng 8	Tháng 9	Tháng 10	Tháng 11	Tháng 12	Hàng năm
1	184,6	185,2	185,0	184,5	184,3	184,0
2	189,5	189,9	190,9	191,0	190,3	188,9
3	196,4	198,8	199,2	197,6	196,8	195,3
4	203,9	202,9	201,8	201,5	201,8	201,6
5	207,917	208,490	208,936	210,177	210,036	207,342
6	219,086	218,783	216,573	212,425	210,228	215,303
7	215,834	215,969	216,177	216,330	215,949	214,537
8	218,312	218,439	218,711	218,803	219,179	218,056
9	226,545	226,889	226,421	226,230	225,672	224,939
10	230,379	231,407	231,317	230,221	229,601	229,594

Lời giải



Hình 2.10

Bài tập vận dụng 2.13

Bảng sau đây là một phần của tập dữ liệu từ www.worldbank.org. Sử dụng bảng để xây dựng đồ thị chuỗi thời gian cho lượng khí thải CO₂ của Hoa Kỳ.

Bảng 2.22

Lượng khí thải CO ₂			
Năm	Ukraine	Vương quốc Anh	Hoa Kỳ
1	352.259	540.640	5.681.664
2	343.121	540.409	5.790.761
3	339.029	541.990	5.826.394
4	327.797	542.045	5.737.615
5	328.357	528.631	5.828.697
6	323.657	522.247	5.656.839
7	272.176	474.579	5.299.563

Công dụng của Đồ thị chuỗi thời gian

Đồ thị chuỗi thời gian là công cụ quan trọng trong các ứng dụng khác nhau của thống kê. Khi ghi lại các giá trị của cùng một biến trong một khoảng thời gian dài, đôi khi khó phân biệt bất kỳ xu hướng hoặc mẫu hình nào. Tuy nhiên, một khi các điểm dữ liệu tương tự được hiển thị bằng đồ thị, một số đặc điểm sẽ nổi bật. Đồ thị chuỗi thời gian giúp dễ phát hiện tính xu hướng.

Làm thế nào để KHÔNG lừa dối bằng thống kê

Điều quan trọng chúng ta cần nhớ là chúng ta phát triển nhiều phương pháp khác nhau để trình bày dữ liệu là để thúc đẩy hiểu biết sâu sắc về đối tượng mà các quan sát đại diện. Chúng ta muốn "cảm nhận" dữ liệu. Có phải các quan sát đều rất giống nhau hay chúng phân tán trên một khoảng rộng giá trị; chúng có tập trung ở một đầu của phổ giá trị hay được phân bố đều. Chúng ta đang cố gắng có được hình ảnh trực quan về dữ liệu ở dạng số. Tiếp ngay sau đây, chúng ta sẽ phát triển phép đo toán học chính thức cho dữ liệu, nhưng trình bày đồ họa trực quan có thể nói lên nhiều điều. Nó cũng có thể, không may, nói nhiều điều gây mất tập trung, gây nhầm lẫn và đơn giản là cho ấn tượng sai từ hình ảnh trực quan. Nhiều năm trước Darrell Huff đã viết cuốn sách *How to Lie with Statistics*. Nó đã qua 25 lần tái bản và bán được hơn một triệu rưỡi bản. Trong cuốn sách, Darren Huff đưa ra quan điểm gay gắt và sử dụng nhiều ví dụ thực tế để cho thấy sự lừa dối. Ông muốn làm cho mọi người nhận thức được sự lừa dối như vậy, nhưng có lẽ quan trọng hơn là giáo dục để những người khác không mắc phải những lỗi tương tự một cách vô tình.

Một lần nữa, mục tiêu của đồ thị là làm sáng tỏ bằng cách sử dụng hình ảnh để kể câu chuyện về dữ liệu. Biểu đồ tròn có một số vấn đề phổ biến khi được sử dụng để truyền đạt

thông điệp của dữ liệu. Biểu đồ với quá nhiều mảnh sẽ làm choáng ngợp người đọc. Có lẽ 5 hoặc 6 mảnh (phân loại) là đủ để đưa ra ý tưởng về tầm quan trọng tương đối của mỗi mảnh (phân loại). Xét cho cùng, đây là mục tiêu của biểu đồ tròn, cho thấy tập con nào quan trọng nhất trong những tập con. Nếu có quá nhiều mảnh thì có lẽ thay cách tiếp cận sẽ tốt hơn, hoặc có lẽ một số tập con nên được hợp nhất thành một (loại "khác"). Biểu đồ tròn không thể hiển thị những thay đổi theo thời gian, mặc dù chúng ta thấy nhiều người cố gắng làm điều này trong thực tế. Trong một số tài liệu tài chính, biểu đồ tròn thường được dùng để hiển thị các cấu phần của nguồn thu ngân sách: thuế thu nhập, thuế giá trị gia tăng, thuế xe cơ giới, v.v.. Các thông tin hữu ích này có thể được thể hiện tốt với biểu đồ tròn. Lỗi là khi dữ liệu hai năm được đặt cạnh nhau. Vì tổng doanh thu thay đổi từ năm này sang năm khác, nhưng kích thước của bánh tròn được cố định, không có thông tin thực sự nào được cung cấp và kích thước tương đối của mỗi mảnh bánh tròn không thể được so sánh một cách có ý nghĩa.

Đồ thị histogram có thể rất hữu ích trong việc hiểu dữ liệu. Nếu được trình bày đúng cách, chúng có thể là một cách trực quan, nhanh chóng trình bày xác suất của các phân loại khác nhau thông qua so sánh trực quan, đơn giản độ lớn tương đối của các phân loại. Ở đây lỗi có thể xảy, có chủ ý hay không, là biến đổi chiều rộng của các phân loại (thanh). Điều này tất nhiên làm cho việc so sánh với các phân loại khác nhau trở nên không thể. Mở rộng chiều rộng của một phân loại "tôn cao" không thỏa đáng tầm quan trọng của phân loại có chiều rộng mở rộng, vì nó có diện tích lớn hơn, và do đó "nói" một cách trực quan rằng phân loại đó có xác suất xảy ra cao hơn.

Đồ thị chuỗi thời gian có lẽ là loại đồ thị bị lạm dụng nhiều nhất. Không thể trình bày đồ thị của một biến theo trên các trục có tỷ lệ thay đổi một phần trên trang giấy, bất kể chiều dọc hoặc chiều ngang. Chẳng hạn khung thời gian được thay đổi từ năm sang tháng. Điều này có thể là để tiết kiệm không gian hoặc vì dữ liệu hàng tháng không có sẵn cho những năm đầu. Trong cả hai trường hợp điều này làm trộn lẫn việc trình bày và hủy hoại ý nghĩa của đồ thị. Nếu điều này không được thực hiện để cố ý làm rối người đọc, thì nó chắc chắn là việc lừa dối hoặc cầu thả.

Thay đổi đơn vị đo lường của trục có thể giúp giảm nhẹ hoặc cường điệu sự biến động. Nếu bạn muốn cho thấy những biến động lớn, thì hãy đo biến bằng đơn vị nhỏ (sử dụng đơn vị xu thay vì hàng nghìn đô la). Và tất nhiên để tiếp tục gian lận, hãy chắc chắn rằng trục không bắt đầu từ điểm (0, 0). Nếu đồ thị bắt đầu từ (0, 0) thì việc trục đồ thị bị thao túng sẽ dễ bị nhận thấy.

Có lẽ bạn có một khách hàng lo lắng về sự biến động của danh mục đầu tư do bạn quản lý. Một cách dễ dàng để trình bày dữ liệu là sử dụng các khoảng thời gian dài trên đồ thị chuỗi thời gian. Sử dụng đơn vị tháng hoặc tốt hơn, quý thay vì dữ liệu hàng ngày hoặc hàng tuần. Nếu điều đó không làm giảm tính biến động thì hãy trải rộng trục thời gian so với trục tỷ lệ hoàn vốn (hoặc định giá danh mục đầu tư). Nếu bạn muốn cho thấy sự tăng trưởng "nhẹ nhàng" ẩn tượng, thì hãy thu nhỏ trục thời gian. Bất kỳ tăng trưởng tích cực nào sẽ thành tốc độ tăng trưởng "cao" một cách trực quan. Lưu ý rằng nếu tăng trưởng là âm thì thủ thuật này sẽ cho thấy danh mục đầu tư đang sụp đổ với tốc độ ẩn tượng.

Một lần nữa, mục tiêu của thống kê mô tả là truyền đạt hình ảnh có ý nghĩa, có thể kể câu chuyện về dữ liệu. Thao túng có chủ ý là gian lận và trong trường hợp tệ nhất là phi đạo đức, ngay cả khi có ý tốt, việc mắc phải những loại lỗi này sẽ dẫn đến nhầm lẫn về phân tích.

2.2 Các số đo vị trí của dữ liệu

Các số đo vị trí phổ biến là **tứ phân vị** (quartiles) và **phân vị** (percentile).

Các tứ phân vị, là số phân vị đặc biệt, chia một tập dữ liệu đã sắp xếp theo trật tự tăng dần thành bốn phần bằng nhau. Một tập dữ liệu có 3 số tứ phân vị, được ký hiệu là Q1, Q2, và Q3.

- Khoảng một phần tư dữ liệu rơi vào hoặc dưới tứ phân vị thứ nhất Q1.
- Khoảng một nửa dữ liệu rơi vào hoặc dưới tứ phân vị thứ hai Q2.
- Khoảng ba phần tư dữ liệu rơi vào hoặc dưới tứ phân vị thứ ba Q3.

Tương tự, phân vị chia một tập dữ liệu thành 100 phần bằng nhau.

Để tính tứ phân vị và phân vị, dữ liệu phải được sắp xếp từ nhỏ nhất đến lớn nhất. Tứ phân vị chia dữ liệu đã sắp xếp thành các phần tư. Phân vị chia dữ liệu đã sắp xếp thành các phần trăm. Đạt điểm ở phân vị thứ 90 trong một kỳ thi không có nghĩa là bạn nhận được điểm 90% trong bài kiểm tra. Nó chỉ có nghĩa là 90% số điểm kiểm tra bằng hoặc thấp hơn điểm của bạn và 10% số điểm kiểm tra bằng hoặc cao hơn điểm kiểm tra của bạn.

Phân vị hữu ích để so sánh các giá trị. Vì lý do này, các trường đại học và cao đẳng sử dụng phân vị rộng rãi. Một trường hợp mà các trường cao đẳng và đại học sử dụng phân vị là khi kết quả SAT được sử dụng để xác định điểm xét tuyển. Ví dụ, giả sử Đại học Duke công bố chấp nhận điểm SAT ở hoặc trên phân vị thứ 75; nghĩa là ứng viên cần có điểm SAT ít nhất 1220.

Thường thì người ta sử dụng phân vị với các quần thể rất lớn. Do đó, nếu bạn nói rằng 90% điểm kiểm tra thấp hơn điểm của bạn (không phải bằng hoặc thấp hơn), thì điều đó cũng chấp nhận được vì việc loại bỏ một điểm dữ liệu cụ thể sẽ không ảnh hưởng đáng kể kết quả so sánh.

Số trung vị (median) là một số đo "tâm" của dữ liệu. Bạn có thể nghĩ về số trung vị như "giá trị giữa"; nó không nhất thiết phải là một trong các giá trị quan sát được. Trung vị chỉ là một con số chia tập dữ liệu (đã sắp xếp) thành hai nửa. Một nửa là các điểm bằng hoặc nhỏ hơn số trung vị, và một nửa chứa các trị bằng hoặc lớn hơn. Ví dụ, hãy xem xét dữ liệu sau.

1; 11,5; 6; 7,2; 4; 8; 9; 10; 6,8; 8,3; 2; 2; 10; 1

Sắp xếp từ nhỏ nhất đến lớn nhất:

1; 1; 2; 2; 4; 6; 6,8; 7,2; 8; 8,3; 9; 10; 10; 11,5

Vì có 14 quan sát, số trung vị nằm giữa giá trị thứ bảy, 6,8, và giá trị thứ tám, 7,2. Để tìm số trung vị, cộng hai giá trị lại với nhau và chia cho hai.

$$6,8 + 7,2 = 14$$

$$14 \div 2 = 7$$

Số trung vị là 7. Một nửa các điểm dữ liệu nhỏ hơn 7 và một nửa các điểm dữ liệu lớn hơn 7.

Tứ phân vị là những số tách dữ liệu thành các phần tư. Tứ phân vị có thể là hoặc không phải là một phần của dữ liệu. Để tìm tứ phân vị, trước tiên hãy tìm số trung vị (hoặc tứ phân vị thứ hai). Tứ phân vị thứ nhất, Q_1 , là giá trị giữa của nửa dưới của dữ liệu, và tứ phân vị thứ ba, Q_3 , là giá trị giữa, hoặc số trung vị, của nửa trên của dữ liệu. Để hiểu ý tưởng, hãy xem xét cùng một tập dữ liệu:

1; 1; 2; 2; 4; 6; 6,8; 7,2; 8; 8,3; 9; 10; 10; 11,5

Số trung vị hoặc **tứ phân vị thứ hai** (2nd quartile) là 7. Nửa dưới của dữ liệu là 1, 1, 2, 2, 4, 6, 6,8. Giá trị giữa của nửa dưới là 2.

1; 1; 2; 2; 4; 6; 6,8

Số 2, là một phần của dữ liệu, là **tứ phân vị thứ nhất** (1st quartile). Một phần tư của toàn bộ tập các giá trị bằng hoặc nhỏ hơn 2 và ba phần tư của các giá trị lớn hơn 2.

Nửa trên của dữ liệu là 7,2, 8, 8,3, 9, 10, 10, 11,5. Giá trị giữa của nửa trên là 9.

Tứ phân vị thứ ba (3rd quartile), Q_3 , là 9. Ba phần tư (75%) của tập dữ liệu đã sắp xếp nhỏ hơn 9. Một phần tư (25%) của tập dữ liệu đã sắp xếp lớn hơn 9. Tứ phân vị thứ ba là một phần của tập dữ liệu trong ví dụ này.

Khoảng tứ phân vị (interquartile range) là chỉ số cho thấy độ phân tán của nửa giữa hoặc 50% giữa của dữ liệu. Nó là hiệu giữa tứ phân vị thứ ba (Q_3) và tứ phân vị thứ nhất (Q_1).

$$IQR = Q_3 - Q_1$$

IQR có thể giúp xác định các **điểm ngoại vi** (outlier) tiềm tàng. **Một giá trị được nghi ngờ là giá trị ngoại vi tiềm tàng nếu nó nhỏ hơn ($Q_1 - 1,5IQR$) hoặc lớn hơn ($Q_3 + 1,5IQR$).** Các giá trị ngoại vi tiềm tàng luôn đòi hỏi điều tra thêm.

LƯU Ý

Giá trị ngoại vi tiềm tàng là một điểm dữ liệu khác biệt đáng kể so với các điểm dữ liệu khác. Những điểm dữ liệu đặc biệt này có thể là lỗi hoặc một loại bất thường nào đó hoặc chúng có thể là chìa khóa để hiểu dữ liệu.

Ví dụ 2.14

Vấn đề

Đối với 13 giá bất động sản sau đây, tính IQR và xác định xem có giá nào là giá trị ngoại vi tiềm tàng không. Giá tính bằng đô la.

389.950; 230.500; 158.000; 479.000; 639.000; 114.950; 5.500.000; 387.000; 659.000; 529.000; 575.000; 488.800; 1.095.000

Lời giải

Sắp xếp dữ liệu từ nhỏ nhất đến lớn nhất.

114.950; 158.000; 230.500; 387.000; 389.950; 479.000; 488.800; 529.000; 575.000; 639.000; 659.000; 1.095.000; 5.500.000

$$M = 488.800$$

$$Q1 = \frac{230.500 + 387.000}{2} = 308.750$$

$$Q3 = \frac{639.000 + 659.000}{2} = 649.000$$

$$IQR = 649.000 - 308.750 = 340.250$$

$$(1,5)(IQR) = (1,5)(340.250) = 510.375$$

$$Q1 - (1,5)(IQR) = 308.750 - 510.375 = -201.625$$

$$Q3 + (1,5)(IQR) = 649.000 + 510.375 = 1.159.375$$

Không có giá nhà nào nhỏ hơn -201.625. Tuy nhiên, 5.500.000 lớn hơn 1.159.375. Do đó, 5.500.000 là một **giá trị ngoại vi** tiềm tàng.

Bài tập vận dụng 2.14

Đối với 11 mức lương sau đây, tính *IQR* và xác định xem có mức lương nào là giá trị ngoại vi không. Mức lương tính bằng đô la.

\$33.000; \$64.500; \$28.000; \$54.000; \$72.000; \$68.500; \$69.000; \$42.000; \$54.000; \$120.000; \$40.500

Ví dụ 2.15

Vấn đề

Điểm kiểm tra cho lớp thống kê ban ngày là:

99; 56; 78; 55,5; 32; 90; 80; 81; 56; 59; 45; 77; 84,5; 84; 70; 72; 68; 32; 79; 90

Điểm kiểm tra cho lớp thống kê buổi tối là:

98; 78; 68; 83; 81; 89; 88; 76; 65; 45; 98; 90; 80; 84,5; 85; 79; 78; 98; 90; 79; 81; 25,5

Đối với hai tập dữ liệu, tìm những điều sau:

- Khoảng tứ phân vị. So sánh hai khoảng tứ phân vị.
- Bất kỳ giá trị ngoại vi nào trong một trong hai tập.

Lời giải

Tóm tắt năm số cho lớp ban ngày và buổi tối là

Bảng 2.23

	Cực tiểu	Q_1	Số trung vị	Q_3	Cực đại
Ban ngày	32	56	74,5	82,5	99
Ban tối	25,5	78	81	89	98

a. IQR cho nhóm ban ngày là $Q3 - Q1 = 82,5 - 56 = 26,5$

IQR cho nhóm buổi tối là $Q3 - Q1 = 89 - 78 = 11$

Khoảng tứ phân vị (độ trải rộng hoặc biến thiên) cho lớp ban ngày lớn hơn IQR lớp ban tối. Điều này cho thấy điểm kiểm tra lớp ban ngày sẽ nhiều biến thiên.

b. Giá trị ngoại vi lớp ban ngày được tìm thấy bằng cách sử dụng quy tắc IQR nhân 1,5. Vậy,

$$Q1 - IQR(1,5) = 56 - 26,5(1,5) = 16,25$$

$$Q3 + IQR(1,5) = 82,5 + 26,5(1,5) = 122,25$$

Vì giá trị cực tiểu và cực đại cho lớp ban ngày lớn hơn 16,25 và nhỏ hơn 122,25 nên không có giá trị ngoại vi.

Giá trị ngoại vi lớp ban tối được tính như sau:

$$Q1 - IQR(1,5) = 78 - 11(1,5) = 61,5$$

$$Q3 + IQR(1,5) = 89 + 11(1,5) = 105,5$$

Đối với lớp này, bất kỳ điểm kiểm tra nào nhỏ hơn 61,5 là một giá trị ngoại vi. Do đó, điểm 45 và 25,5 là giá trị ngoại vi. Vì không có điểm kiểm tra nào lớn hơn 105,5, không có giá trị ngoại vi đầu trên.

Bài tập vận dụng 2.15

Tìm khoảng tứ phân vị cho hai tập dữ liệu sau đây và so sánh chúng.

Điểm kiểm tra cho Lớp A

69; 96; 81; 79; 65; 76; 83; 99; 89; 67; 90; 77; 85; 98; 66; 91; 77; 69; 80; 94

Điểm kiểm tra cho Lớp B

90; 72; 80; 92; 90; 97; 92; 75; 79; 68; 70; 80; 99; 95; 78; 73; 71; 68; 95; 100

Ví dụ 2.16

Năm mươi sinh viên lớp thống kê được hỏi họ ngủ bao nhiêu tiếng mỗi đêm học (làm tròn đến giờ gần nhất). Kết quả là:

Bảng 2.24

Số giờ ngủ mỗi đêm học (giờ)	Tần suất	Tần suất tương đối	Tần suất tương đối tích lũy
4	2	0,04	0,04
5	5	0,10	0,14
6	7	0,14	0,28
7	12	0,24	0,52
8	14	0,28	0,80
9	7	0,14	0,94
10	3	0,06	1,00

Tìm phân vị thứ 28. Chú ý 0,28 trong cột "tần suất tương đối tích lũy". 28% của 50 điểm dữ liệu là 14 điểm. Có 14 điểm có giá nhỏ hơn phân vị thứ 28. Chúng bao gồm hai số 4, năm số 5, và bảy số 6. Phân vị thứ 28 nằm giữa số 6 cuối cùng và số 7 đầu tiên. **Phân vị thứ 28 là 6.5.**

Tìm số trung vị. Nhìn lại cột "tần suất tương đối tích lũy" và tìm 0,52. Số trung vị là phân vị thứ 50 hoặc tứ phân vị thứ hai. 50% của 50 là 25. Có 25 giá trị nhỏ hơn số trung vị. Chúng bao gồm hai số 4, năm số 5, bảy số 6, và 11 số 7 (của 12 số 7 trong tập dữ liệu). Số trung vị hoặc phân vị thứ 50 nằm giữa giá trị thứ 25 (số 7), và thứ 26 (số 7). **Số trung vị là 7.**

Tìm tứ phân vị thứ ba. Tứ phân vị thứ ba giống như phân vị thứ 75. Bạn có thể "nhìn lướt" câu trả lời này. Nếu bạn nhìn vào cột "tần suất tương đối tích lũy", bạn tìm thấy 0.52 và 0.80. Khi bạn có tất cả các số 4, 5, 6, 7 bạn có 52% dữ liệu. Khi bạn bao gồm thêm tất cả các số 8, bạn có 80% dữ liệu. **Phân vị thứ 75, do đó, phải là số 8.** Cách khác để nhìn vào vấn đề là tìm 75% của 50, là 37.5, và làm tròn lên 38. Tứ phân vị thứ ba, Q_3 , là giá trị thứ 38, là số 8. Bạn có thể kiểm tra câu trả lời này bằng cách đếm các giá trị. (Có 37 giá trị dưới tứ phân vị thứ ba và 12 giá trị trên.)

Bài tập vận dụng 2.16

Bốn mươi tài xế xe buýt được hỏi họ dành bao nhiêu giờ mỗi ngày để chạy tuyến đường của họ (làm tròn đến giờ gần nhất). Tìm phân vị thứ 65.

Bảng 2.24

Số thời gian dành cho tuyến đường (giờ)	Tần suất	Tần suất Tương đối	Tần suất Tương đối Tích lũy
2	12	0,30	0,30
3	14	0,35	0,65
4	10	0,25	0,90
5	4	0,10	1,00

Công thức tìm phân vị thứ k

Nếu bạn làm một chút nghiên cứu, bạn sẽ tìm thấy nhiều công thức khác nhau để tính phân vị thứ k . Đây là một trong số chúng.

k = phân vị thứ k . Nó có thể là hoặc không phải là một phần của dữ liệu.

i = chỉ số (thứ hạng hoặc vị trí của một giá trị dữ liệu)

n = tổng số các điểm dữ liệu hay quan sát

- Sắp xếp dữ liệu từ nhỏ nhất đến lớn nhất.
- Tính $i = \frac{k}{100}(n + 1)$
- Nếu i là một số nguyên, thì phân vị thứ k là giá trị dữ liệu ở vị trí thứ i trong tập dữ liệu đã sắp xếp.
- Nếu i không phải là một số nguyên, thì làm tròn i lên và làm tròn i xuống đến các số nguyên gần nhất. Tính trung bình của hai giá trị dữ liệu ở hai vị trí này trong tập dữ liệu đã sắp xếp. Công thức này dễ hiểu hơn qua một ví dụ.

Ví dụ 2.17

Vấn đề

Liệt kê dưới đây là 29 độ tuổi của các nam diễn viên chính giành giải Oscar *theo thứ tự từ nhỏ nhất đến lớn nhất*.

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

a. Tìm phân vị thứ 70.

b. Tìm phân vị thứ 83.

Lời giải

a. $k = 70$

i = chỉ số

$n = 29$

$i = \frac{k}{100}(n + 1) = \left(\frac{70}{100}\right)(29 + 1) = 21$. 21 là một số nguyên, và giá trị dữ liệu ở vị trí thứ 21 trong tập dữ liệu đã sắp xếp là 64. Phân vị thứ 70 là 64 tuổi.

b. $k = \text{phân vị thứ } 83$

i = chỉ số

$n = 29$

$i = \frac{k}{100}(n + 1) = (\frac{83}{100})(29 + 1) = 24,9$, không phải là một số nguyên. Làm tròn xuống 24 và lên 25. Tuổi ở vị trí thứ 24 là 71 và tuổi ở vị trí thứ 25 là 72. Trung bình 71 và 72. Phân vị thứ 83 là 71,5 tuổi.

Bài tập vận dụng 2.17

Liệt kê dưới đây là 29 độ tuổi của các nam diễn viên chính giành giải Oscar theo thứ tự từ nhỏ nhất đến lớn nhất.

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

Tính phân vị thứ 20 và phân vị thứ 55.

LƯU Ý

Sử dụng Excel và R để tìm phân vị thứ k

Excel: Excel cung cấp một số hàm để tìm phân vị thứ k. Hàm PERCENTILE.EXC(data, k) cho kết quả tương tự công thức trên. k từ 0 đến 1, k không bao gồm 0 và 1. Tương tự, để tìm tứ phân vị, sử dụng hàm QUARTILE.EXC(data, k), k = 1, 2, 3.

R/stats package: hàm quantile() cung cấp nhiều công thức tìm phân vị k. Hàm quantile(data, k, type=6) cho kết quả tương tự công thức trên. k từ 0 đến 1. Để tìm tứ phân vị, cho k = 0.25; 0.5; 0.75.

Ví dụ 2.18

Vấn đề

Sử dụng Bảng 2.23:

- Tìm phân vị thứ 80.
- Tìm phân vị thứ 90.
- Tìm tứ phân vị thứ nhất. Tên khác của tứ phân vị thứ nhất là gì?

Lời giải

Sử dụng dữ liệu từ bảng tần suất, chúng ta có:

- Chú ý có 50 giá trị dữ liệu trong bảng, vậy $n = 50$.

Tính chỉ số i như sau:

$$i = \frac{80}{100}(50 + 1) = 40.8$$

Vì $i = 40,8$, tính trung bình của giá trị dữ liệu thứ 40 và thứ 41. Giá trị dữ liệu thứ 40 là 8, giá trị dữ liệu thứ 41 là 9, và trung bình của hai giá trị dữ liệu này là 8,5. Do đó, phân vị thứ 80 là 8,5.

- b. Tính chỉ số i như sau:

$$i = \frac{90}{100}(50 + 1) = 45,9$$

Vì $i = 45,9$, tính trung bình của giá trị dữ liệu thứ 45 và thứ 46. Giá trị dữ liệu thứ 45 là 9, giá trị dữ liệu thứ 46 là 9, và trung bình của hai giá trị dữ liệu này là 9. Do đó, phân vị thứ 90 là 9.

- c. Tên khác của tứ phân vị thứ nhất là phân vị thứ 25. Tiến hành tính phân vị thứ 25:

Tính chỉ số i như sau:

$$i = \frac{25}{100}(50 + 1) = 12.75$$

Vì $i = 12,75$, tính trung bình của giá trị dữ liệu thứ 12 và thứ 13. Giá trị dữ liệu thứ 12 là 6, giá trị dữ liệu thứ 13 là 6, và trung bình của hai giá trị dữ liệu này là 6. Do đó, tứ phân vị thứ nhất là 6.

Bài tập vận dụng 2.18

Tham khảo Bảng 2.24. Tìm tứ phân vị thứ ba. Tên khác của tứ phân vị thứ ba là gì?

Bài tập nhóm

Giảng viên hoặc một thành viên của lớp sẽ hỏi mọi người trong lớp họ sở hữu bao nhiêu áo thun. Trả lời các câu hỏi sau:

1. Có bao nhiêu sinh viên được khảo sát?
2. Bạn đã thực hiện loại lấy mẫu nào?
3. Xây dựng hai đồ thị histogram khác nhau. Đối với mỗi cái, giá trị bắt đầu = _____ giá trị kết thúc = _____.
4. Tìm số trung vị, tứ phân vị thứ nhất và tứ phân vị thứ ba.
5. Xây dựng bảng dữ liệu để tìm:
 - a. phân vị thứ 10
 - b. phân vị thứ 70
 - c. phần trăm sinh viên sở hữu ít hơn bốn thun.

Công thức tìm hạng phân vị (percentile rank) của một giá trị trong tập dữ liệu

- Sắp xếp dữ liệu từ nhỏ nhất đến lớn nhất.
- x = số lượng điểm dữ liệu đếm từ dưới danh sách dữ liệu lên nhưng không bao gồm giá trị dữ liệu mà bạn muốn tìm phân vị.
- y = số lượng điểm dữ liệu bằng giá trị dữ liệu mà bạn muốn tìm phân vị.
- n = tổng số lượng điểm dữ liệu.
- Tính $\frac{x+0,5y}{n}(100)$. Sau đó làm tròn đến số nguyên gần nhất.

Ví dụ 2.19

Vấn đề

Liệt kê dưới đây là 29 độ tuổi của các nam diễn viên chính giành giải Oscar theo thứ tự từ nhỏ nhất đến lớn nhất.

18; 21; 22; 25; 26; 27; 29; 30; 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

- Tìm hạng phân vị cho 58.
- Tìm hạng phân vị cho 25.

Lời giải

- Đếm từ dưới danh sách, có 18 giá trị dữ liệu nhỏ hơn 58. Có một giá trị 58.

$$x = 18 \text{ và } y = 1 \cdot \frac{x+0,5y}{n}(100) = \frac{18+0,5(1)}{29}(100) = 63,80. 58 \text{ là phân vị thứ } 64.$$

- Đếm từ dưới danh sách, có ba giá trị dữ liệu nhỏ hơn 25. Có một giá trị 25.

$$x = 3 \text{ và } y = 1 \cdot \frac{x+0,5y}{n}(100) = \frac{3+0,5(1)}{29}(100) = 12,07. 25 \text{ là phân vị thứ } 12.$$

Bài tập vận dụng 2.19

Liệt kê dưới đây là 30 độ tuổi của các nam diễn viên chính giành giải Oscar theo thứ tự từ nhỏ nhất đến lớn nhất.

18; 21; 22; 25; 26; 27; 29; 30; 31, 31; 33; 36; 37; 41; 42; 47; 52; 55; 57; 58; 62; 64; 67; 69; 71; 72; 73; 74; 76; 77

Tìm phân vị cho 47 và 31.

LƯU Ý

Sử dụng Excel để tìm hạng phân vị

Excel: Excel cung cấp một số hàm để tìm hạng phân vị của giá trị. Hàm PERCENTRANK.EXC(data, x) với x là giá trị cần tìm hạng. Công thức của Excel sẽ cho kết quả hơi khác với công thức trên.

Diễn giải về phân vị, tứ phân vị và trung vị

Phân vị chỉ ra vị trí tương đối của một điểm dữ liệu khi dữ liệu được sắp xếp theo thứ tự từ nhỏ nhất đến lớn nhất. $p\%$ các điểm dữ liệu nhỏ hơn hoặc bằng phân vị thứ p . Ví dụ, 15% các điểm dữ liệu nhỏ hơn hoặc bằng phân vị thứ 15.

- Phân vị thấp luôn tương ứng với giá trị dữ liệu thấp hơn.
- Phân vị cao luôn tương ứng với giá trị dữ liệu cao hơn.

Phân vị có thể tương ứng hoặc không tương ứng với một nhận định về việc nó "tốt" hay "xấu". Việc diễn giải rằng một phân vị nhất định "tốt" hay "xấu" phụ thuộc vào bối cảnh của tình huống mà dữ liệu áp dụng. Trong một số tình huống, phân vị thấp sẽ được coi là "tốt"; trong bối cảnh khác, phân vị cao có thể được coi là "tốt". Trong nhiều tình huống, không có nhận định nào về trị.

Hiểu cách diễn giải đúng về phân vị rất quan trọng không chỉ khi mô tả dữ liệu, mà còn khi tính xác suất trong các chương sau của giáo trình này.

LƯU Ý

Khi trình bày diễn giải về một phân vị trong bối cảnh của dữ liệu được cung cấp, diễn giải nên chứa thông tin sau.

- thông tin về bối cảnh của tình huống đang được xem xét
- giá trị dữ liệu (giá trị của biến) đại diện cho phân vị
- tỷ lệ phần trăm cá thể hoặc đối tượng có giá trị dữ liệu dưới phân vị
- tỷ lệ phần trăm cá thể hoặc đối tượng có giá trị dữ liệu trên phân vị.

Ví dụ 2.20

Vấn đề

Trong một kỳ thi toán có tính giờ, tứ phân vị thứ nhất cho thời gian hoàn thành bài thi là 35 phút. Giải thích tứ phân vị thứ nhất trong bối cảnh của tình huống này.

Lời giải

- Hai mươi lăm phần trăm học sinh hoàn thành bài thi trong 35 phút hoặc ít hơn.
- Bảy mươi lăm phần trăm học sinh hoàn thành bài thi trong 35 phút hoặc nhiều hơn.
- Phân vị thấp có thể được coi là tốt, vì hoàn thành nhanh hơn trong bài thi có thời gian là mong muốn. (Nếu bạn mất quá nhiều thời gian, bạn có thể không thể hoàn thành.)

Bài tập vận dụng 2.20

Đối với cuộc thi chạy 100 mét, tứ phân vị thứ ba cho thời gian hoàn thành cuộc thi là 11,5 giây. Giải thích tứ phân vị thứ ba trong bối cảnh của tình huống.

Ví dụ 2.21

Vấn đề

Trong một bài kiểm tra toán 20 câu, phân vị thứ 70 cho số câu trả lời đúng là 16. Giải thích phân vị thứ 70 trong bối cảnh của tình huống này.

Lời giải

- 70% sinh viên trả lời đúng 16 câu hỏi hoặc ít hơn.
- 30% sinh viên trả lời đúng 16 câu hỏi hoặc nhiều hơn.
- Phân vị cao hơn được coi là tốt vì việc trả lời đúng nhiều câu hỏi là điều kỳ vọng.

Bài tập vận dụng 2.21

Trong một bài tập viết 60 điểm, phân vị thứ 80 cho số điểm kiếm được là 49. Giải thích phân vị thứ 80 trong bối cảnh của tình huống này.

Ví dụ 2.22

Vấn đề

Tại một trường cao đẳng cộng đồng, người ta thấy rằng phân vị thứ 30 của số tín chỉ mà học sinh đăng ký là bảy tín chỉ. Giải thích phân vị thứ 30 trong bối cảnh của tình huống này.

Lời giải

- 30% sinh viên đăng ký bảy tín chỉ hoặc ít hơn.
- 70% sinh viên đăng ký bảy tín chỉ hoặc nhiều hơn.
- Trong ví dụ này, không đánh giá giá trị "tốt" hay "xấu" liên quan đến phân vị cao hơn hoặc thấp hơn. Sinh viên học tại trường cao đẳng cộng đồng vì nhiều lý do và nhu cầu khác nhau, và khối lượng học tập của họ thay đổi tùy theo nhu cầu cá nhân.

Bài tập vận dụng 2.22

Trong một mùa giải, phân vị thứ 40 cho điểm ghi được mỗi cầu thủ trong một trò chơi là 8. Giải thích phân vị thứ 40 trong bối cảnh của tình huống này.

Ví dụ 2.23*Vấn đề*

Trường Trung học Sharpe đang xin tài trợ để trang bị thêm thiết bị tập cho phòng gym. Hiệu trưởng khảo sát ẩn danh 15 học sinh để xác định bao nhiêu phút một ngày các học sinh dành cho tập thể dục. Kết quả từ 15 học sinh như sau.

0 phút; 40 phút; 60 phút; 30 phút; 60 phút

10 phút; 45 phút; 30 phút; 300 phút; 90 phút;

30 phút; 120 phút; 60 phút; 0 phút; 20 phút

Xác định năm giá trị sau.

Min = 0

$Q_1 = 20$

Med = 40

$Q_3 = 60$

Max = 300

Nếu bạn là hiệu trưởng, bạn có biện minh được cho việc mua thiết bị tập mới không? Vì 75% học sinh tập thể dục 60 phút hoặc ít hơn hàng ngày, và vì IQR là 40 phút ($60 - 20 = 40$), chúng ta biết rằng một nửa số học sinh được khảo sát tập thể dục từ 20 phút đến 60 phút hàng ngày. Đây có vẻ là thời lượng hợp lý dành cho tập thể dục, vì vậy số liệu biện minh cho việc hiệu trưởng muốn mua thiết bị mới.

Tuy nhiên, hiệu trưởng cần cẩn thận. Giá trị 300 có vẻ là một giá trị ngoại vi tiềm tàng.

$Q_3 + 1,5(IQR) = 60 + (1,5)(40) = 120$.

Giá trị 300 lớn hơn 120 nên nó là một giá trị ngoại vi tiềm tàng. Nếu chúng ta xóa nó và tính lại 5 số đo, chúng ta có những giá trị sau:

Min = 0

$Q_1 = 20$

$Q_3 = 60$

Max = 120

Chúng ta vẫn có 75% học sinh tập thể dục 60 phút hoặc ít hơn hàng ngày và một nửa số học sinh tập thể dục từ 20 đến 60 phút một ngày. Tuy nhiên, 15 học sinh là một mẫu nhỏ và hiệu trưởng nên khảo sát thêm học sinh để chắc chắn về kết quả khảo sát của mình.

Bài tập vận dụng 2.23

Một giảng viên môn thống kê nghiên cứu thời lượng mà sinh viên dành để làm dự án cuối kỳ của môn học. Giảng viên muốn sinh viên dành khoảng 3 đến 4 giờ cho dự án như là thời lượng điển hình. Giảng viên thu thập dữ liệu về số giờ dành cho dự án cuối kỳ từ mẫu ngẫu nhiên 10 sinh viên. Kết quả thu được như sau:

2 giờ; 3 giờ; 5 giờ; 3 giờ; 4 giờ;

4 giờ; 3 giờ; 11 giờ; 3 giờ; 2 giờ.

- Xác định năm giá trị sau: Min, Q1, Med, Q3, Max.
- Giảng viên có nên điều chỉnh dự án cuối kỳ hay để như cũ?

LƯU Ý

Sử dụng Excel, R để tìm trị nhỏ nhất và lớn nhất

Excel: cung cấp hàm MIN(data), MAX(data) để tìm trị nhỏ nhất và lớn nhất từ dữ liệu.

R: cung cấp hàm min(data), max(data) để tìm trị nhỏ nhất và lớn nhất từ dữ liệu.

2.3 Các số đo tâm của dữ liệu

"Tâm" của một tập dữ liệu cũng là một cách mô tả vị trí. Hai số đo "tâm" dữ liệu phổ biến nhất là số **trung bình** (mean, còn gọi là trung bình) và số **trung vị** (median). Để tính **cân nặng trung bình** của 50 người, ta cộng 50 số đo cân nặng lại với nhau và chia cho 50. Để tìm **cân nặng trung vị** của 50 người, ta sắp xếp dữ liệu và tìm số chia dữ liệu thành hai phần bằng nhau. Số trung vị thường là thước đo tâm tốt hơn khi dữ liệu có các giá trị cực đoan hoặc giá trị ngoại vi vì nó không bị ảnh hưởng bởi các giá trị của các điểm ngoại vi. Số bình quân là thước đo tâm phổ biến nhất.

LƯU Ý

Các từ "bình quân" và "trung bình" thường được sử dụng thay thế cho nhau. Đây là việc phổ biến. "Bình quân số học" là tham số; "trung bình" là thông số thống kê. Về mặt toán học, "bình quân số học" còn là moment thứ nhất của hàm phân phối. Tuy nhiên, trong thực tế giữa những người không chuyên thống kê, "trung bình" thường được chấp nhận cho "bình quân số học".

Khi mỗi giá trị trong tập dữ liệu không phải là duy nhất, số bình quân có thể được tính bằng cách nhân mỗi giá trị riêng biệt với tần suất của nó và sau đó chia tổng cho tổng số giá trị dữ liệu. Chữ cái được sử dụng để đại diện cho **số bình quân mẫu** (sample mean) là một $\bar{x}$ có thanh ở trên (đọc là "x gạch", "x bar"): $\bar{x}$.

Chữ cái Hy Lạp μ (đọc là "miu") đại diện cho **số bình quân quần thể** (population mean). Một trong những yêu cầu để **số bình quân mẫu** là ước lượng tốt của **số bình quân quần thể** là mẫu được lấy phải thực sự ngẫu nhiên.

Để thấy rằng cả hai cách tính số bình quân đều giống nhau, hãy xem xét mẫu:

1; 1; 1; 2; 2; 3; 4; 4; 4; 4; 4

$$\bar{x} = \frac{1 + 1 + 1 + 2 + 2 + 3 + 4 + 4 + 4 + 4 + 4}{11} = 2,7$$

$$\bar{x} = \frac{3(1) + 2(2) + 1(3) + 5(4)}{11} = 2,7$$

Trong phép tính thứ hai, các tần suất là 3, 2, 1 và 5.

Bạn có thể nhanh chóng tìm vị trí của số trung vị bằng cách sử dụng biểu thức $\frac{n+1}{2}$.

Chữ cái n là tổng số lượng điểm dữ liệu trong mẫu. Nếu n là một số lẻ, số trung vị là giá trị giữa của dữ liệu đã sắp xếp (sắp xếp từ nhỏ nhất đến lớn nhất). Nếu n là một số chẵn, số trung vị bằng hai giá trị giữa cộng lại với nhau và chia cho hai sau khi dữ liệu đã được sắp xếp. Ví dụ, nếu tổng số lượng điểm dữ liệu là 97, thì $\frac{n+1}{2} = \frac{97+1}{2} = 49$. Số trung vị là điểm dữ liệu thứ 49 trong dữ liệu đã sắp xếp. Nếu tổng số giá trị dữ liệu là 100, thì $\frac{n+1}{2} = \frac{100+1}{2} = 50,5$. Số trung vị nằm giữa 2 điểm dữ liệu thứ 50 và thứ 51. Vị trí của số trung vị và giá trị của số trung vị **không** giống nhau. Chữ cái viết hoa M thường được sử dụng để đại diện cho số trung vị. Ví dụ tiếp theo minh họa vị trí của số trung vị và giá trị của số trung vị.

Ví dụ 2.24

Vấn đề

Một quản trị viên bệnh viện theo dõi tuổi (tính bằng năm) của bệnh nhân đến phòng cấp cứu trong khoảng thời gian một tuần (dữ liệu được sắp xếp từ nhỏ nhất đến lớn nhất):
3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27;
27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;

Tính số bình quân và số trung vị.

Lời giải

Phép tính cho số bình quân là:

$$\bar{x} = \frac{[3+4+(8)(2)+10+11+12+13+14+(15)(2)+(16)(2)+\dots+35+37+40+(44)(2)+47]}{40} = 23,6$$

Để tìm số trung vị, M , trước tiên sử dụng công thức cho vị trí. Vị trí là:
 $\frac{n+1}{2} = \frac{40+1}{2} = 20,5$

Bắt đầu từ giá trị nhỏ nhất, số trung vị nằm giữa giá trị thứ 20 và thứ 21 (hai số 24):
3; 4; 8; 8; 10; 11; 12; 13; 14; 15; 15; 16; 16; 17; 17; 18; 21; 22; 22; 24; 24; 25; 26; 26; 27;
27; 29; 29; 31; 32; 33; 33; 34; 34; 35; 37; 40; 44; 44; 47;

$$M = \frac{24 + 24}{2} = 24$$

Bài tập vận dụng 2.24

Dữ liệu sau đây cho thấy số tháng bệnh nhân thường chờ đợi trong danh sách cấy ghép trước khi được phẫu thuật. Dữ liệu được sắp xếp từ nhỏ nhất đến lớn nhất. Tính số bình quân và số trung vị.

3; 4; 5; 7; 7; 7; 7; 8; 8; 9; 9; 10; 10; 10; 10; 10; 10; 11; 12; 12; 13; 14; 14; 15; 15; 17; 17; 18; 19; 19; 19; 21; 21; 22; 22; 23; 24; 24; 24; 24

Ví dụ 2.25

Vấn đề

Giả sử trong một thị trấn nhỏ có 50 người, một người kiếm được \$5.000.000 mỗi năm và 49 người còn lại mỗi người kiếm được \$30.000. Số đo nào tốt hơn cho "tâm" của dữ liệu: số bình quân hay số trung vị?

Lời giải

$$\bar{x} = \frac{5.000.000 + 49(30.000)}{50} = 129.400$$

$$M = 30.000$$

(Có 49 người kiếm được \$30.000 và một người kiếm được \$5.000.000.)

Số trung vị là số đo tốt hơn cho "tâm" so với số bình quân vì 49 trong số các giá trị là 30.000 và một là 5.000.000. Số 5.000.000 là một giá trị ngoại vi. Số 30.000 cho chúng ta cảm nhận tốt hơn về tâm của dữ liệu.

Bài tập vận dụng 2.25

Trong một mẫu 60 hộ gia đình, một ngôi nhà trị giá \$2.500.000. 29 ngôi nhà trị giá \$280.000, và tất cả các ngôi nhà khác trị giá \$315.000. Số đo nào tốt hơn cho "tâm" của dữ liệu: số bình quân hay số trung vị?

Một thước đo khác của trung tâm là **yếu vị** (mode). Yếu vị là giá trị xuất hiện thường xuyên nhất trong tập dữ liệu. Có thể có nhiều hơn một yếu vị trong một tập dữ liệu miễn là những giá trị đó có cùng tần suất và tần suất đó là cao nhất. Một tập dữ liệu có hai yếu vị được gọi là song yếu vị.

Ví dụ 2.26

Điểm kiểm tra môn thống kê của 20 sinh viên như sau:

50; 53; 59; 59; 63; 63; 72; 72; 72; 72; 72; 76; 78; 81; 83; 84; 84; 84; 90; 93

Vấn đề

Tìm yếu vị.

Lời giải

Điểm xuất hiện thường xuyên nhất là 72, xuất hiện năm lần. Yếu vị = 72.

Bài tập vận dụng 2.26

Số sách được mượn từ thư viện của 25 sinh viên như sau:

0; 0; 0; 1; 2; 3; 3; 4; 4; 5; 5; 7; 7; 7; 7; 8; 8; 8; 9; 10; 10; 11; 11; 12; 12

Tìm yếu vị.

Ví dụ 2.27

Ta có điểm kiểm tra môn bất động sản là 430, 430, 480, 480, 495. Tập dữ liệu là song yếu vị vì điểm 430 và 480 mỗi điểm xuất hiện hai lần.

Khi nào yếu vị là thước đo tốt nhất của "tâm"? Hãy xem xét một chương trình giảm cân quảng cáo rằng trung bình số cân giảm là 6 pound trong tuần đầu tiên của chương trình. Tuy nhiên, số yếu vị chỉ ra rằng hầu hết mọi người chỉ giảm 2 pound trong tuần đầu tiên; như vậy chương trình có vẻ kém hấp dẫn hơn.

LƯU Ý

Yếu vị có thể được tính cho dữ liệu định tính cũng như cho dữ liệu định lượng. Ví dụ, nếu tập dữ liệu là: đỏ, đỏ, đỏ, xanh lá, xanh lá, vàng, tím, đen, xanh dương, yếu vị là đỏ.

Bài tập vận dụng 2.27

Ta có dữ liệu điểm tín dụng là 680, 680, 700, 720, 720. Tập dữ liệu là song yếu vị vì điểm 680 và 720 mỗi điểm xuất hiện hai lần. Hãy xem xét thu nhập hàng năm của công nhân tại một nhà máy. Yếu vị là \$25.000 và xuất hiện 150 lần trong số 301 điểm dữ liệu. Số trung vị là \$50.000 và số bình quân là \$47.500. Số đo nào sẽ là tốt nhất cho "tâm" dữ liệu?

LƯU Ý

Sử dụng Excel, R để tìm bình quân, trung vị, yếu vị.

Excel: cung cấp hàm AVERAGE(data), MEDIAN(data), MODE(data) để tìm bình quân, trung vị và yếu vị.

R: cung cấp hàm mean(data), median(data) để tìm bình quân và trung vị.

Tính số bình quân của bảng tần suất nhóm

Khi dữ liệu bị nhóm (grouped data), bạn không biết cụ thể trị của từng điểm dữ liệu (chúng ta chỉ biết các khoảng và tần suất khoảng). Do đó, bạn không thể tính số bình quân chính xác cho tập dữ liệu. Điều chúng ta phải làm là ước lượng số bình quân thực sự bằng cách tính số bình quân của bảng tần suất. **Bảng tần suất** (frequency table) là một cách trình bày dữ liệu trong đó dữ liệu nhóm được hiển thị cùng với các tần suất tương ứng. Để tính số bình quân từ bảng tần suất nhóm, chúng ta có thể áp dụng định nghĩa cơ bản của số bình quân: $\text{số bình quân} = \frac{\text{tổng dữ liệu}}{\text{số lượng điểm dữ liệu}}$. Chúng ta chỉ cần sửa đổi định nghĩa để phù hợp với các hạn chế của bảng tần suất.

Vì chúng ta không biết cụ thể giá trị các điểm dữ liệu, thay vào đó chúng ta có thể tìm trung điểm của mỗi khoảng. **Trung điểm** (midpoint) là $\frac{\text{biên dưới} + \text{biên trên}}{2}$. Bây giờ chúng ta có thể sửa đổi định nghĩa số bình quân bình quân bảng tần suất $= \frac{\sum fm}{\sum f}$ trong đó f = tần suất của khoảng và m = trung điểm của khoảng.

Ví dụ 2.28

Vấn đề

Sau đây là một bảng tần suất điểm bài kiểm tra cuối môn thống kê của giáo sư Blount. Tìm ước lượng tốt nhất cho số bình quân điểm kiểm tra.

Bảng 2.26

Khoảng điểm	Số sinh viên
50 - 56,5	1
56,5 - 62,5	0
62,5 - 68,5	4
68,5 - 74,5	4
74,5 - 80,5	2
80,5 - 86,5	3
86,5 - 92,5	4
92,5 - 98,5	1

Lời giải

Tìm trung điểm cho tất cả các khoảng

Bảng 2.27

Khoảng điểm	Điểm giữa
50-56,5	53,25
56,5-62,5	59,5
62,5-68,5	65,5
68,5-74,5	71,5
74,5-80,5	77,5
80,5-86,5	83,5
86,5-92,5	89,5
92,5-98,5	95,5

Tính tổng của tích của mỗi tần suất khoảng và trung điểm $\sum f m$

$$53,25(1) + 59,5(0) + 65,5(4) + 71,5(4) + 77,5(2) + 83,5(3) + 89,5(4) + 95,5(1) = 1460,25$$

$$\mu = \frac{\sum fm}{\sum f} = \frac{1460,25}{19} = 76,86$$

Bài tập vận dụng 2.28

Một nhà nghiên cứu đã tiến hành một nghiên cứu về tác động mà chơi trò chơi điện tử lên trí nhớ. Như một phần của nghiên cứu, họ tập hợp dữ liệu sau:

Bảng 2.28

Số giờ thanh thiếu niên dành cho trò chơi điện tử	Số thanh thiếu niên
0-3,5	3
3,5-7,5	7
7,5-11,5	12
11,5-15,5	7
15,5-19,5	9

Ước lượng tốt nhất cho số giờ bình quân dành để chơi trò chơi điện tử là gì?

2.4 Ký hiệu sigma và tính trung bình số học

Công thức cho trung bình quần thể

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

Công thức cho trung bình mẫu

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

Phần này nhằm nhắc lại những kiến thức bạn đã từng học và hẳn đã nói rằng "tôi chắc chắn sẽ không bao giờ cần đến điều này".

Đây là các công thức cho trung bình quần thể và trung bình mẫu. Chữ cái Hy Lạp μ (miu) là ký hiệu cho trung bình quần thể và $\bar{x}$ là ký hiệu cho trung bình mẫu. Cả hai công thức đều có một ký hiệu toán học chỉ ra cách thực hiện phép tính. Nó được gọi là ký hiệu sigma vì ký hiệu này là chữ cái Hy Lạp hoa sigma: Σ . Giống như tất cả các ký hiệu toán học, dấu + báo hiệu phép cộng và dấu $\times$ báo hiệu phép nhân, nó hướng dẫn chúng ta thực hiện phép toán. Những ký hiệu này được gọi là các toán tử toán học. Ký hiệu Σ nhắc chúng ta cộng một danh sách cụ thể các số.

Giả sử có một mẫu động vật từ trạm cứu hộ động vật địa phương và chúng ta quan tâm đến độ tuổi trung bình của chúng. Nếu chúng ta liệt kê từng giá trị, hoặc quan sát, trong một cột, bạn có thể đánh số thứ tự cho từng giá trị. Số đầu tiên sẽ là số 1 và số thứ hai là số 2, v.v.

Bảng 2.29

Động vật	Tuổi
1	9
2	1
3	8,5
4	10,5
5	10
6	8,5
7	12
8	8
9	1
10	9,5

Mỗi quan sát đại diện cho một động vật cụ thể trong mẫu. Purr là động vật số một và là một con mèo 9 tuổi, Toto là động vật số hai và là một con chó con 1 tuổi, v.v.

Để tính trung bình, theo công thức chúng ta cộng tất cả các số này, trong trường hợp này là tuổi, rồi chia tổng cho 10, tổng số động vật trong mẫu.

Động vật số 1, mèo Purr, được ký hiệu là x_1 , động vật số 2, Toto, được ký hiệu là x_2 , v.v. cho đến Dundee là động vật số 10 và được ký hiệu là x_{10} .

Chữ i trong công thức chỉ ra những quan sát nào cần cộng lại với nhau. Trong trường hợp này là x_1 đến x_{10} nghĩa là tất cả chúng. Nếu chúng ta sử dụng ký hiệu chỉ số, $i = 1$ và vì nó là một mẫu nên chúng ta sử dụng n nhỏ ở trên cùng của Σ sẽ là 10.

Độ lệch chuẩn cũng cần cùng một ký hiệu toán học nên sẽ cần thiết khi nhớ trước kiến thức này.

Tổng tuổi được tính là 78 và chia cho 10 cho chúng ta tuổi trung bình mẫu là 7,8 năm.

2.5 Trung bình nhân

Trung bình (cộng), trung vị và yếu vị đều là các phép đo "trung tâm", "trung bình" của dữ liệu. Tất cả đều, theo một cách riêng, cố gắng đo lường điểm "bình thường" hay "điển hình" của dữ liệu. Trong trường hợp của trung bình cộng, điều này được giải quyết bằng cách tìm giá trị mà từ đó tất cả các điểm đều có khoảng cách tuyến tính bằng nhau. Chúng ta có thể tưởng tượng rằng tất cả các giá trị dữ liệu được kết hợp thông qua phép cộng và sau đó được phân phối lại cho từng điểm dữ liệu với lượng bằng nhau sao cho tổng tất cả các giá trị vẫn bằng tổng ban đầu.

Trung bình nhân (geometric mean) không phân phối lại tổng các giá trị mà phân phối lại tích của việc nhân tất cả các giá trị riêng lẻ và sau đó phân phối lại chúng thành các phần bằng nhau sao cho tích chung vẫn giữ nguyên. Điều này có thể được thấy từ công thức cho trung bình nhân, $\tilde{x}$: (Đọc là "x ngã")

$$\tilde{x} = \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}} = \sqrt[n]{x_1 \cdot x_2 \dots x_n} = (x_1 \cdot x_2 \dots x_n)^{\frac{1}{n}}$$

trong đó π là một ký hiệu toán học khác, yêu cầu chúng ta nhân tất cả các số x_i theo cách tương tự như ký hiệu Hy Lạp sigma nhắc cộng tất cả các số x_i . Lưu ý, số mũ phân số là căn bậc n của số đó.

Trung bình nhân trả lời câu hỏi: "nếu tất cả các đại lượng có cùng một giá trị, thì giá trị đó phải là bao nhiêu để đạt được cùng một tích?" Trung bình nhân có tên gọi như vậy bởi thực tế là khi được phân phối lại theo cách này, các cạnh tạo thành một hình học mà tất cả các cạnh có cùng độ dài. Để thấy điều này, hãy lấy ví dụ về các số 10, 51,2 và 8. Trung bình nhân là tích của việc nhân ba số này với nhau (4.096) và lấy căn bậc ba vì có ba số mà tích này được phân phối giữa chúng. Do đó, trung bình nhân của ba số này là 16. Điều này mô tả một hình lập phương 16x16x16 và có thể tích là 4.096 đơn vị.

Trung bình nhân có ý nghĩa trong kinh tế học và tài chính để tính toán tăng trưởng: tăng trưởng của thị trường, của đầu tư, dân số và các biến số khác mà người ta quan tâm. Hãy tưởng tượng rằng hợp 4.096 đơn vị của chúng ta (có thể là đô la) là giá trị của một khoản đầu tư sau ba năm và tỷ suất sinh lời của khoản đầu tư tính theo phần trăm là ba số trong ví dụ. Trung bình nhân sẽ cung cấp cho chúng ta câu trả lời cho câu hỏi, tỷ suất sinh lời

trung bình là bao nhiêu: 16 phần trăm. Trung bình cộng của ba số này là 23,6 phần trăm. Lý do cho sự khác biệt này, 16 so với 23,6, là trung bình cộng tính cộng dồn và do đó không tính đến lãi sinh ra từ lãi, lãi kép, được phát sinh trong quá trình tăng trưởng đầu tư. Vấn đề tương tự khi hỏi về tỷ lệ tăng trưởng trung bình của dân số hoặc doanh số hoặc thâm nhập thị trường, v.v., khi biết các tỷ lệ tăng trưởng hàng năm. Công thức cho tỷ suất sinh lời trung bình nhân, hoặc bất kỳ tỷ lệ tăng trưởng nào khác, là:

$$r_s = (x_1 \cdot x_2 \dots x_n)^{\frac{1}{n}} - 1$$

Thao tác với công thức trung bình nhân cũng có thể cung cấp một phép tính tỷ lệ tăng trưởng trung bình giữa hai kỳ chỉ cần biết giá trị ban đầu và giá trị cuối và số kỳ. Công thức sau đây cung cấp thông tin này:

$$\left(\frac{a_n}{a_0}\right)^{\frac{1}{n}} = \tilde{x}$$

Cuối cùng, chúng ta lưu ý rằng công thức cho trung bình nhân yêu cầu tất cả các số phải dương, lớn hơn không. Lý do hiển nhiên là việc tính căn của số âm không có nghĩa trong các ứng dụng thực tế, chỉ tồn tại trong lý thuyết toán học thuần túy. Có những cách để giải quyết vấn đề này, tuy nhiên trong trường hợp tỷ lệ hoàn vốn và bài toán tăng trưởng đơn, ta có thể chuyển các giá trị âm thành các giá trị dương tương đương có ý nghĩa. Hãy tưởng tượng rằng lợi nhuận hàng năm trong ba năm qua là -1,2%, -8% và 2,02. Sử dụng công thức lợi nhuận hàng năm tương đương là 0,98; 0,92 và 1,02 cho phép sử dụng công thức nhân của trung bình nhân là +1,0167. Trừ 1 từ giá trị này cho chúng ta trung bình nhân là +1,67% như một tỷ lệ tăng trưởng dân số ròng (hoặc lợi nhuận tài chính). Từ ví dụ này, chúng ta có thể thấy rằng trung bình nhân cung cấp cho chúng ta với công thức này để tính toán trung bình nhân của một chuỗi tỷ lệ lợi nhuận hàng năm:

$$r_s = \tilde{x} - 1$$

trong đó r_s là tỷ lệ lợi nhuận trung bình và $\tilde{x}$ là trung bình nhân của các lợi nhuận trong một số giai đoạn thời gian nào đó. Lưu ý rằng độ dài của mỗi giai đoạn thời gian phải giống nhau.

Theo nguyên tắc chung, ta nên chuyển đổi các giá trị phần trăm sang dạng số nhân thập phân tương đương. Cần lưu ý rằng khi làm việc với phần trăm, trung bình nhân của các giá trị phần trăm khác với trung bình nhân của các số nhân thập phân tương đương, và chính trung bình nhân của các số nhân thập phân tương đương mới là giá trị có ý nghĩa.

LƯU Ý

Sử dụng Excel để tìm trung bình nhân.

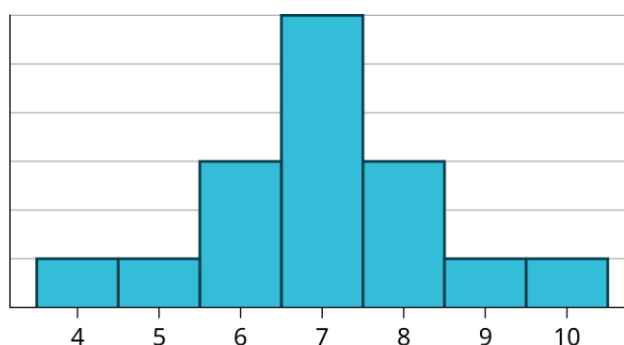
Excel: cung cấp hàm GEOMEAN(data).

2.6 Độ lệch và bình quân, trung vị, và yếu vị

Hãy xem xét tập dữ liệu sau:

4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 7; 8; 8; 8; 9; 10

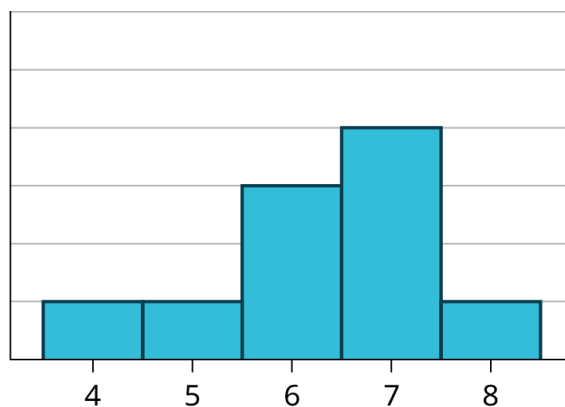
Tập dữ liệu này có thể được biểu diễn bằng đồ thị histogram sau. Mỗi khoảng có chiều rộng là một, và mỗi giá trị nằm ở giữa một khoảng.



Hình 2.11

Đồ thị histogram (được hiển thị trong Hình 2.11) thể hiện một phân phối dữ liệu **đối xứng** (symmetrical). Một phân phối là đối xứng nếu có thể vẽ một đường thẳng đứng tại một điểm nào đó trong đồ thị histogram sao cho hình dạng bên trái và bên phải của đường thẳng đứng là hình ảnh phản chiếu của nhau. Đối với tập dữ liệu này bình quân, trung vị và yếu vị đều là 7. **Trong một phân phối hoàn toàn đối xứng** (symmetrical distribution), **bình quân và trung vị là giống nhau**. Ví dụ này có một yếu vị (đơn yếu vị, unimodal), và yếu vị bằng bình quân và trung vị. Trong một phân phối đối xứng có hai yếu vị (song yếu vị, bimodal), hai yếu vị sẽ khác với bình quân và trung vị.

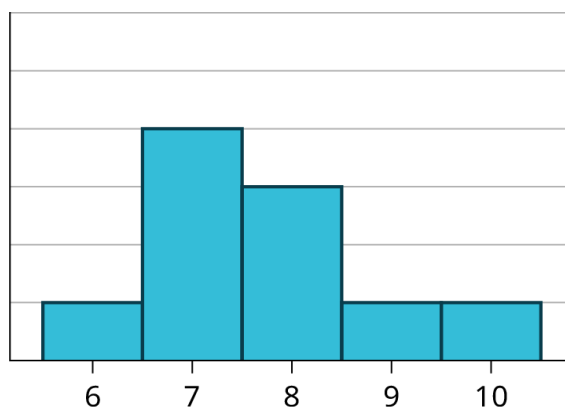
Đồ thị histogram cho dữ liệu: 4; 5; 6; 6; 6; 7; 7; 7; 7; 7; 8 (được hiển thị trong Hình 2.12) không đối xứng. Phía bên phải có vẻ "bị cắt bớt" so với phía bên trái. Phân phối loại này được gọi là **lệch trái** (skewed to the left) vì nó được kéo ra về phía trái. Chúng ta có thể đo lường chính xác độ lệch của một phân phối giống như cách chúng ta đo lường toán học giá trị trung tâm của dữ liệu hay độ "phân tán" chung của nó. Công thức toán học tính độ lệch là: $a_3 = \frac{\sum (x_i - \bar{x})^3}{ns^3}$. Độ lệch càng xa số 0 thì mức độ lệch của phân phối càng lớn. Nếu độ lệch có giá trị âm thì phân phối lệch trái như trong Hình 2.12. Thước đo độ lệch dương biểu thị sự lệch phải như trong Hình 2.13.



Hình 2.12

Trong trường hợp này, bình quân là 6,3, trung vị là 6,5 và yếu vị là 7. **Chú ý rằng bình quân nhỏ hơn trung vị, và cả hai đều nhỏ hơn yếu vị.** Cả bình quân và trung vị đều phản ánh độ lệch, nhưng số bình quân phản ánh nhiều hơn.

Đồ thị histogram cho dữ liệu: 6; 7; 7; 7; 7; 8; 8; 8; 8; 9; 10 (Hình 2.13), cũng không đối xứng. Nó **lệch phải** (skewed to the right).



Hình 2.13

Bình quân là 7,7, trung vị là 7,5 và yếu vị là 7. Trong ba thông số thống kê, **bình quân là lớn nhất, trong khi yếu vị là nhỏ nhất.** Một lần nữa, bình quân phản ánh độ lệch nhiều nhất.

Bình quân bị ảnh hưởng bởi các giá trị ngoại vi, nhưng các trị ngoại vi lại không ảnh hưởng đến trung vị. Do đó, khi phân phối dữ liệu lệch trái, bình quân thường nhỏ hơn trung vị. Khi phân phối lệch phải, bình quân thường lớn hơn trung vị. Trong các phân phối đối xứng, chúng ta kỳ vọng bình quân và trung vị xấp xỉ bằng nhau về giá trị. Đây là một sự liên kết quan trọng giữa hình dạng của phân phối và mối quan hệ của bình quân và trung vị. Tuy nhiên, điều này không đúng cho mọi tập dữ liệu. Các ngoại lệ phổ biến nhất xảy ra đối với các tập dữ liệu rời rạc.

$$a_3 = \sum \frac{(x_i - \bar{x})^3}{ns^3}$$

trong đó s là độ lệch chuẩn mẫu của dữ liệu, x_i và $\bar{x}$ là trung bình cộng và n là cỡ mẫu.

Trong toán học, bình quân số học được biết đến như là mômen bậc nhất của phân phối. Mômen bậc hai mà chúng ta thấy là phương sai, và độ lệch là mômen bậc ba. Phương sai đo lường bình phương của các độ lệch giữa dữ liệu và giá trị trung bình, trong khi độ lệch đo lường lập phương của các độ lệch giữa dữ liệu và giá trị trung bình. Mặc dù phương sai không bao giờ có giá trị âm, nhưng độ lệch có thể âm hoặc dương, và điều này giúp chúng ta xác định dữ liệu lệch về bên trái hay bên phải. Độ lệch của phân phối chuẩn bằng 0, và bất kỳ dữ liệu đối xứng nào cũng có độ lệch gần bằng 0. Giá trị âm của độ lệch cho biết dữ liệu lệch trái, còn giá trị dương cho biết dữ liệu lệch phải. Lệch trái có nghĩa là đuôi bên trái dài hơn so với đuôi bên phải. Tương tự, lệch phải có nghĩa là đuôi bên phải dài hơn so với đuôi bên trái. Độ lệch đặc trưng cho mức độ bất đối xứng của phân phối quanh giá trị trung bình. Trong khi trung bình và độ lệch chuẩn là những đại lượng có thứ nguyên (đây là lý do tại sao chúng ta lấy căn bậc hai của phương sai), nghĩa là có cùng đơn vị với các đại lượng X_i được đo, thì độ lệch được định nghĩa theo cách làm cho nó không có thứ nguyên. Nó là một con số thuần túy chỉ đặc trưng cho hình dạng của phân phối. Giá trị dương của độ lệch biểu thị phân phối có đuôi bất đối xứng kéo dài về phía giá trị X lớn hơn, còn giá trị âm biểu thị phân phối có đuôi kéo dài về phía giá trị X nhỏ hơn. Độ lệch bằng 0 cho thấy phân phối đối xứng.

Độ lệch và tính đối xứng sẽ trở nên quan trọng khi chúng ta nghiên cứu về phân phối xác suất trong các chương tiếp theo.

LƯU Ý

Sử dụng Excel để tìm độ lệch và yếu vị.

Excel: cung cấp hàm SKEW(data) và MODE(data) để tìm độ lệch và yếu vị.

2.7 Số đo độ phân tán của dữ liệu

Một đặc điểm quan trọng của bất kỳ tập dữ liệu nào là sự biến thiên trong dữ liệu. Trong một số tập dữ liệu, các giá trị dữ liệu tập trung gần bình quân; trong các tập dữ liệu khác, các giá trị dữ liệu trải rộng ra xa hơn từ bình quân. Số đo biến thiên, hay độ phân tán, phổ biến nhất là độ lệch chuẩn. **Độ lệch chuẩn** (standard deviation) là một số đo lường mức độ các giá trị dữ liệu cách xa bình quân của chúng.

Độ lệch chuẩn

- cung cấp một số đo về tổng lượng biến thiên trong một tập dữ liệu, và
- có thể được sử dụng để xác định liệu một giá trị dữ liệu cụ thể gần hay xa bình quân.

Độ lệch chuẩn cung cấp số đo của tổng biến thiên trong một tập dữ liệu

Độ lệch chuẩn luôn dương hoặc bằng không. Độ lệch chuẩn nhỏ khi tất cả dữ liệu tập trung gần bình quân, thể hiện ít biến thiên hoặc độ phân tán nhỏ. Độ lệch chuẩn lớn hơn khi các giá trị dữ liệu phân tán cách xa bình quân, thể hiện nhiều biến thiên hơn.

Giả sử chúng ta đang nghiên cứu thời gian khách hàng xếp hàng chờ đợi thanh toán tại siêu thị A và siêu thị B. Thời gian chờ đợi trung bình tại cả hai siêu thị là 5 phút. Tại siêu

thị A, độ lệch chuẩn cho thời gian chờ đợi là 2 phút; tại siêu thị B độ lệch chuẩn cho thời gian chờ đợi là 4 phút.

Vì siêu thị B có độ lệch chuẩn cao hơn, chúng ta biết rằng có nhiều biến thiên hơn trong thời gian chờ đợi tại siêu thị B. Nhìn chung, thời gian chờ đợi tại siêu thị B trải rộng ra xa trung bình; thời gian chờ đợi tại siêu thị A tập trung xung quanh trung bình.

Tính toán độ lệch chuẩn

Nếu x là một số, thì sự khác biệt " x - số bình quân" được gọi là **độ lệch** (deviation) của nó. Một tập dữ liệu có bao nhiêu phần tử thì sẽ có bấy nhiêu độ lệch. Các độ lệch được sử dụng để tính độ lệch chuẩn. Nếu các số thuộc về một quần thể, độ lệch được ký hiệu là $x - \mu$. Đối với mẫu, độ lệch được ký hiệu là $x - \bar{x}$.

Thủ tục tính toán độ lệch chuẩn phụ thuộc vào việc dữ liệu là toàn bộ quần thể hay là một mẫu. Các phép tính tương tự, nhưng không giống hệt. Do đó ký hiệu độ lệch chuẩn cho tổng thể và mẫu khác nhau. Chữ la-tinh s đại diện cho độ lệch chuẩn mẫu và chữ cái Hy Lạp σ (sigma, chữ thường) đại diện cho độ lệch chuẩn quần thể. Nếu mẫu có những đặc điểm giống như quần thể, thì s sẽ là một ước lượng tốt của σ .

Để tính toán độ lệch chuẩn, chúng ta cần tính toán phương sai trước. **Phương sai** (variance) là **trung bình của bình phương các độ lệch** (các giá trị $x - \bar{x}$ cho một mẫu, hoặc các giá trị $x - \mu$ cho một tổng thể). Ký hiệu σ^2 đại diện cho phương sai quần thể; độ lệch chuẩn quần thể σ là căn bậc hai của phương sai quần thể. Ký hiệu s^2 đại diện cho phương sai mẫu; độ lệch chuẩn mẫu s là căn bậc hai của phương sai mẫu. Bạn có thể xem độ lệch chuẩn như là một trung bình đặc biệt của các độ lệch. Về mặt lý thuyết, phương sai là mômen bậc hai của phân phối, hay còn gọi là mômen bậc nhất của bình quân. Cần nhớ rằng bình quân chính là mômen bậc nhất của phân phối.

Nếu dữ liệu đến từ một **khảo sát toàn bộ quần thể**, không phải từ một mẫu, khi chúng ta tính trung bình của các bình phương độ lệch để tìm phương sai, chúng ta chia cho N , số lượng phần tử trong tổng thể. Nếu dữ liệu từ một **mẫu** thay vì là một tổng thể, khi chúng ta tính trung bình của các bình phương độ lệch, chúng ta chia cho $n - 1$, ít hơn một so với số phần tử trong mẫu.

Công thức cho độ lệch chuẩn mẫu

- $s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}}$ hoặc $s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n-1}}$ hoặc $s = \sqrt{\frac{(\sum_{i=1}^n x_i^2) - n\bar{x}^2}{n-1}}$
- Đối với độ lệch chuẩn mẫu, mẫu số là $n - 1$, tức là kích thước mẫu TRỪ 1.

Công thức cho độ lệch chuẩn quần thể

- $\sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}}$ hoặc $\sigma = \sqrt{\frac{\sum f(x - \mu)^2}{N}}$ hoặc $\sigma = \sqrt{\frac{\sum_{i=1}^N x_i^2}{N} - \mu^2}$
- Đối với độ lệch chuẩn quần thể, mẫu số là N , số phần tử trong quần thể.

Trong những công thức này, f đại diện cho tần suất mà một giá trị xuất hiện. Ví dụ, nếu một giá trị xuất hiện một lần, f là 1. Nếu một giá trị xuất hiện ba lần trong tập dữ liệu hoặc tổng thể, f là 3. Hai nhận xét quan trọng về phương sai và độ lệch chuẩn: các độ lệch được

do từ giá trị trung bình và các độ lệch được bình phương. Về nguyên tắc, các độ lệch có thể được đo từ bất kỳ điểm nào, tuy nhiên, chúng ta quan tâm đến việc đo từ trọng tâm của dữ liệu, tức là giá trị "bình thường" hoặc thông thường nhất của quan sát. Sau này chúng ta sẽ cố gắng đo lường tính "bất thường" của một quan sát hoặc một trung bình mẫu; do đó chúng ta cần một thước đo từ giá trị trung bình. Nhận xét thứ hai là các độ lệch được bình phương. Điều này có hai tác dụng: thứ nhất là làm cho tất cả các độ lệch đều dương và thứ hai là thay đổi đơn vị đo từ đơn vị của trung bình và các quan sát ban đầu. Nếu dữ liệu là trọng lượng thì trung bình được đo bằng pound, nhưng phương sai được đo bằng pound bình phương. Một lý do để sử dụng độ lệch chuẩn là để quay trở lại đơn vị đo ban đầu bằng cách lấy căn bậc hai của phương sai. Hơn nữa, khi các độ lệch được bình phương, nó làm tăng mạnh giá trị của chúng. Ví dụ, một độ lệch 10 so với trung bình khi được bình phương là 100, nhưng một độ lệch 100 so với trung bình là 10.000. Điều này làm cho các giá trị ngoại vi có trọng số rất lớn khi tính phương sai.

Các loại biến thiên trong mẫu

Khi nghiên cứu một quần thể, người ta thường sử dụng một mẫu, vì lý do thuận tiện hoặc vì không thể tiếp cận toàn bộ quần thể. Biến thiên là thuật ngữ dùng để mô tả những khác biệt có thể xảy ra trong các kết quả này.

Các loại biến thiên phổ biến bao gồm:

- Biến thiên quan sát hoặc đo lường
- Biến thiên tự nhiên
- Biến thiên gây ra
- Biến thiên mẫu

Dưới đây là một số ví dụ để mô tả từng loại biến thiên.

Ví dụ 1: Biến thiên đo lường

Biến thiên đo lường (measurement variability) xảy ra khi có sự khác biệt về các công cụ được sử dụng để đo lường hoặc giữa những người sử dụng các công cụ đó. Nếu chúng ta thu thập dữ liệu về thời gian một quả bóng rơi từ một độ cao bằng cách cho sinh viên đo thời gian rơi bằng đồng hồ bấm giờ, chúng ta có thể gặp phải biến thiên đo lường nếu hai đồng hồ bấm giờ được sử dụng là của các nhà sản xuất khác nhau. Ví dụ, một đồng hồ đo chính xác đến giây, trong khi đồng hồ kia đo chính xác đến 1/10 giây. Chúng ta cũng có thể gặp phải biến thiên đo lường vì có hai người khác nhau thu thập dữ liệu. Thời gian phản ứng của họ khi nhấn nút trên đồng hồ bấm giờ có thể khác nhau; do đó, các kết quả sẽ thay đổi tương ứng. Sự khác biệt trong kết quả có thể bị ảnh hưởng bởi biến thiên đo lường.

Ví dụ 2: Biến thiên tự nhiên

Biến thiên tự nhiên (natural variability) phát sinh từ những khác biệt xảy ra một cách tự nhiên vì các thành viên của quần thể khác nhau. Ví dụ, nếu chúng ta có hai cây ngô giống hệt nhau và cho cả hai cây cùng một lượng nước và ánh sáng mặt trời, chúng vẫn có thể phát triển với tốc độ khác nhau đơn giản vì chúng là hai cây ngô khác nhau. Sự khác biệt trong kết quả có thể được giải thích bởi biến thiên tự nhiên.

Ví dụ 3: Biến thiên gây ra

Biến thiên gây ra (induced variability) là đối lập của biến thiên tự nhiên; điều này xảy ra vì chúng ta đã tạo ra một yếu tố biến thiên một cách nhân tạo (theo định nghĩa, yếu tố này không có sẵn trong tự nhiên). Ví dụ, chúng ta chia mọi người thành hai nhóm khác nhau để nghiên cứu trí nhớ, và chúng ta tạo ra một biến thiên trong một nhóm bằng cách hạn chế số giờ ngủ của họ. Sự khác biệt trong kết quả có thể bị ảnh hưởng bởi biến thiên gây ra.

Ví dụ 4: Biến thiên mẫu

Biến thiên mẫu (sample variability) xảy ra khi nhiều mẫu ngẫu nhiên được lấy từ cùng một quần thể. Ví dụ, nếu tôi thực hiện bốn cuộc khảo sát với 50 người được chọn ngẫu nhiên từ một quần thể nhất định, sự khác biệt trong kết quả có thể bị ảnh hưởng bởi biến thiên mẫu.

Ví dụ 2.29

Trong một lớp học, giáo viên quan tâm đến tuổi trung bình và độ lệch chuẩn mẫu của tuổi học sinh. Dữ liệu sau đây là tuổi cho một MẪU $n = 20$ học sinh lớp năm. Tuổi được làm tròn đến nửa năm gần nhất:

9; 9,5; 9,5; 10; 10; 10; 10; 10,5; 10,5; 10,5; 10,5; 11; 11; 11; 11; 11; 11; 11,5; 11,5; 11,5;

$$\bar{x} = \frac{9 + 9,5(2) + 10(4) + 10,5(4) + 11(6) + 11,5(3)}{20} = 10,525$$

Tuổi trung bình là 10,53 năm, làm tròn đến hai chữ số.

Phương sai có thể được tính bằng cách sử dụng một bảng. Sau đó độ lệch chuẩn được tính bằng cách lấy căn bậc hai của phương sai. Chúng ta sẽ giải thích các phần của bảng sau khi tính s .

Bảng 2.30

Dữ liệu	Tần suất	Độ lệch	Độ lệch ²	(Tần suất)(Độ lệch ²)
x	f	$(x - \bar{x})$	$(x - \bar{x})^2$	$(f)(x - \bar{x})^2$
9	1	$9 - 10,525 = -1,525$	$(-1,525)^2 = 2,325625$	$1 \times 2,325625 = 2,325625$
9,5	2	$9,5 - 10,525 = -1,025$	$(-1,025)^2 = 1,050625$	$2 \times 1,050625 = 2,101250$
10	4	$10 - 10,525 = -0,525$	$(-0,525)^2 = 0,275625$	$4 \times 0,275625 = 1,1025$
10,5	4	$10,5 - 10,525 = -0,025$	$(-0,025)^2 = 0,000625$	$4 \times 0,000625 = 0,0025$
11	6	$11 - 10,525 = 0,475$	$(0,475)^2 = 0,225625$	$6 \times 0,225625 = 1,35375$
11,5	3	$11,5 - 10,525 = 0,975$	$(0,975)^2 = 0,950625$	$3 \times 0,950625 = 2,851875$
				Tổng là 9,7375

Phương sai mẫu, s^2 , bằng tổng của cột cuối cùng (9,7375) chia cho tổng số giá trị dữ liệu trừ một (20 - 1):

$$s^2 = \frac{9,7375}{20 - 1} = 0,5125$$

Độ lệch chuẩn mẫu s bằng căn bậc hai của phương sai mẫu:

$$s = \sqrt{0,5125} = 0,715891, \text{ được làm tròn đến hai chữ số thập phân, } s = 0,72.$$

Bài tập vận dụng 2.29

Trong một đội bóng chày, tuổi của mỗi cầu thủ như sau:

21; 21; 22; 23; 24; 24; 25; 25; 28; 29; 29; 31; 32; 33; 33; 34; 35; 36; 36; 36; 36; 38; 38; 38; 40

Sử dụng máy tính hoặc máy tính của bạn để tìm số bình quân và độ lệch chuẩn. Sau đó tìm giá trị cao hơn hai độ lệch chuẩn so với bình quân.

LƯU Ý

Sử dụng Excel và để tìm phương sai và độ lệch chuẩn.

Excel: cung cấp hàm VAR.P(data) và STDEV.P(data) để tìm phương sai và độ lệch chuẩn của quần thể; hàm VAR.S(data) và STDEV.S(data) để tìm phương sai và độ lệch chuẩn của mẫu.

R: hàm var(data) và sd(data) để tìm phương sai và độ lệch chuẩn của mẫu.

Giải thích về phép tính độ lệch chuẩn được hiển thị trong bảng

Các độ lệch cho thấy dữ liệu phân tán xung quanh số bình quân như thế nào. So với giá trị dữ liệu 11, giá trị dữ liệu 11,5 xa bình quân hơn như được chỉ ra bởi các độ lệch 0,47 và 0,97. Độ lệch dương xảy ra khi giá trị dữ liệu lớn hơn bình quân, trong khi độ lệch âm xảy ra khi giá trị dữ liệu nhỏ hơn bình quân. Độ lệch là -1,525 cho giá trị dữ liệu 9. **Nếu bạn cộng các độ lệch, tổng luôn bằng không** (Ví dụ 2.29, có $n = 20$ độ lệch.) Vậy bạn không thể đơn giản cộng các độ lệch để có được độ phân tán của dữ liệu. Bằng cách bình phương các độ lệch, bạn làm cho chúng thành số dương, và tổng cũng sẽ dương. Phương sai, do đó, là trung bình bình phương độ lệch. Bằng cách bình phương các độ lệch, chúng ta đang tạo ra một tác động mạnh đối với những quan sát xa giá trị trung bình; những quan sát này có trọng số lớn hơn trong các phép tính phương sai. Điều này có ý nghĩa quan trọng trong thống kê suy luận, nơi phương sai và độ lệch chuẩn đóng vai trò then chốt trong việc xác định các kết luận. Ngay từ bây giờ, chúng ta có thể sử dụng độ lệch chuẩn như một thước đo tính "bất thường". "Bạn làm bài kiểm tra thế nào?" "Tuyệt vời! Cao hơn hai

độ lệch chuẩn so với giá trị trung bình." Như chúng ta sẽ thấy, đây là một điểm kiểm tra tốt một cách bất thường.

Phương sai là một số đo bình phương và không có cùng đơn vị như dữ liệu. Việc lấy căn bậc hai giải quyết vấn đề này. Độ lệch chuẩn đo lường độ phân tán với cùng đơn vị đo như dữ liệu.

Chú ý rằng thay vì chia cho $n = 20$, phép tính chia cho $n - 1 = 20 - 1 = 19$ vì dữ liệu là một mẫu. Đối với phương sai **mẫu**, chúng ta chia cho kích thước mẫu trừ đi một ($n - 1$). Tại sao không chia cho n ? Câu trả lời liên quan đến phương sai quần thể. **Phương sai mẫu là một ước lượng của phương sai quần thể**. Ước lượng này yêu cầu chúng ta phải sử dụng giá trị ước tính của bình quân quần thể thay vì giá trị bình quân quần thể thực tế. Dựa trên cơ sở toán học nằm đằng sau những phép tính này, chia cho $(n - 1)$ cho một ước lượng tốt hơn của phương sai quần thể.

Độ lệch chuẩn, s hoặc σ , hoặc bằng 0 hoặc lớn hơn 0. Sự phân tán của dữ liệu được gọi là "biến thiên". Biến thiên trong dữ liệu phụ thuộc vào phương pháp mà các dữ liệu được thu thập; ví dụ, bằng cách đo lường hoặc bằng cách lấy mẫu ngẫu nhiên. Khi độ lệch chuẩn bằng 0, không có sự phân tán; tức là, tất cả các giá trị dữ liệu bằng nhau. Độ lệch chuẩn nhỏ khi tất cả dữ liệu tập trung gần bình quân, và lớn hơn khi các giá trị dữ liệu thể hiện nhiều biến thiên hơn từ bình quân. Khi độ lệch chuẩn lớn hơn nhiều so với 0, các giá trị dữ liệu phân tán rất rộng quanh bình quân; các giá trị ngoại vi cũng có thể làm s hoặc σ rất lớn.

Ví dụ 2.30

Vấn đề

Sử dụng dữ liệu sau đây (điểm kiểm tra đầu tiên) từ lớp tiền giải tích mùa xuân của Giáo sư Dean:

33; 42; 49; 49; 53; 55; 55; 61; 63; 67; 68; 68; 69; 69; 72; 73; 74; 78; 80; 83; 88; 88; 88; 90; 92; 94; 94; 94; 94; 96; 100

- Tạo một biểu đồ chứa dữ liệu, tần suất, tần suất tương đối và tần suất tương đối tích lũy đến ba chữ số thập phân.
- Tính những điều sau đến một chữ số thập phân:
 - Số trung bình mẫu
 - Độ lệch chuẩn mẫu
 - Số trung vị
 - Tứ phân vị thứ nhất
 - Tứ phân vị thứ ba
 - IQR

Lời giải

- Xem Bảng 2.31
- Số bình quân mẫu = 73,5
 - Độ lệch chuẩn mẫu = 17,9

iii. Số trung vị = 73

iv. Tứ phân vị thứ nhất = 61

v. Tứ phân vị thứ ba = 90

vi. IQR = 90 - 61 = 29

Bảng 2.31

Dữ liệu	Tần suất	Tần suất tương đối	Tần suất tương đối tích lũy
33	1	0,032	0,032
42	1	0,032	0,064
49	2	0,065	0,129
53	1	0,032	0,161
55	2	0,065	0,226
61	1	0,032	0,258
63	1	0,032	0,29
67	1	0,032	0,322
68	2	0,065	0,387
69	2	0,065	0,452
72	1	0,032	0,484
73	1	0,032	0,516
74	1	0,032	0,548
78	1	0,032	0,580
80	1	0,032	0,612
83	1	0,032	0,644
88	3	0,097	0,741
90	1	0,032	0,773
92	1	0,032	0,805
94	4	0,129	0,934
96	1	0,032	0,966
100	1	0,032	0,998 (Tại sao giá trị này không phải 1?)

Bài tập vận dụng 2.30

Dữ liệu sau đây cho thấy các loại thức ăn cho thú cưng khác nhau mà các cửa hàng trong khu vực bán.

6; 6; 6; 6; 7; 7; 7; 7; 7; 8; 9; 9; 9; 9; 10; 10; 10; 10; 10; 11; 11; 11; 11; 12; 12; 12; 12; 12; 12;

Tính số trung bình mẫu và độ lệch chuẩn mẫu đến một chữ số thập phân.

Độ lệch chuẩn của Bảng tần suất theo nhóm

Nhớ rằng đối với dữ liệu nhóm chúng ta không biết các giá trị dữ liệu cá nhân, vì vậy chúng ta không thể mô tả giá trị điển hình của dữ liệu với độ chính xác. Nói cách khác, chúng ta không thể tìm chính xác số bình quân, trung vị hoặc yếu vị. Tuy nhiên, chúng ta có thể xác định ước lượng tốt nhất của các thước đo trung tâm bằng cách tìm trung bình của dữ liệu nhóm với công thức:

$$\text{Trung bình của bảng tần suất} = \frac{\sum fm}{\sum f}$$

trong đó f = tần suất khoảng và m = điểm giữa khoảng.

Cũng như chúng ta không thể tìm số bình quân chính xác, chúng ta cũng không thể tìm độ lệch chuẩn chính xác. Nhớ rằng độ lệch chuẩn mô tả số về độ lệch dự kiến mà một giá trị dữ liệu có từ số bình quân. Nói đơn giản, độ lệch chuẩn cho phép chúng ta so sánh mức độ "bất thường" của dữ liệu cá nhân so với số bình quân.

Ví dụ 2.31

Vấn đề

Tìm độ lệch chuẩn cho dữ liệu trong Bảng 2.32.

Bảng 2.32

Lớp	Tần suất, f	Điểm giữa, m	$f \cdot m$	$f(m - \bar{x})^2$
0-2	1	1	$1 \cdot 1 = 1$	$1(1 - 6,88)^2 = 34,57$
3-5	6	4	$6 \cdot 4 = 24$	$6(4 - 6,88)^2 = 49,77$
6-8	10	7	$10 \cdot 7 = 70$	$10(7 - 6,88)^2 = 0,14$
9-11	7	10	$7 \cdot 10 = 70$	$7(10 - 6,88)^2 = 68,14$
12-14	0	13	$0 \cdot 13 = 0$	$0(13 - 6,88)^2 = 0$
	$n=24$		$\bar{x} = \frac{165}{24} = 6,88$	$s^2 = \frac{152,62}{24 - 1} = 6,64$

Đối với tập dữ liệu này, chúng ta có trung bình $\bar{x} = 6,88$ và độ lệch chuẩn $s_x = 2,58$. Điều này có nghĩa là một giá trị dữ liệu được chọn ngẫu nhiên dự kiến sẽ cách trung bình 2,58 đơn vị. Nếu chúng ta xem xét lớp đầu tiên, chúng ta thấy rằng điểm giữa lớp bằng 1. Giá trị này cách trung bình gần ba độ lệch chuẩn. Trong khi công thức để tính độ lệch chuẩn

không phức tạp, $s_x = \sqrt{\frac{\sum f(m-\bar{x})^2}{n-1}}$ trong đó

s_x = độ lệch chuẩn mẫu, $\bar{x}$ = trung bình mẫu, các phép tính ở đây khá tẻ nhạt. Những phép tính này thường được thực hiện bằng công nghệ.

Bài tập vận dụng 2.31

Tìm độ lệch chuẩn cho dữ liệu từ ví dụ trước

Bảng 2.33

Lớp	Tần suất, f
0-2	1
3-5	6
6-8	10
9-11	7
12-14	0
15-17	2

So sánh giá trị từ các tập dữ liệu khác nhau

Độ lệch chuẩn hữu ích khi so sánh các giá trị dữ liệu đến từ các tập dữ liệu khác nhau. Nếu các tập dữ liệu có số bình quân và độ lệch chuẩn khác nhau, thì việc so sánh các giá trị dữ liệu trực tiếp có thể gây hiểu lầm.

- Đối với mỗi giá trị dữ liệu x , tính có bao nhiêu độ lệch chuẩn cách xa số bình quân của nó.
- Sử dụng công thức: $x = \text{trung bình} + (\text{\#ofSTDEVs})(\text{độ lệch chuẩn})$; tính $\text{\#ofSTDEVs}$.
- $\text{\#ofSTDEVs} = \frac{x - \text{trung bình}}{\text{độ lệch chuẩn}}$
- So sánh kết quả của phép tính này.
- $\text{\#ofSTDEVs}$ thường được gọi là “điểm z ” (z -score); chúng ta có thể sử dụng ký hiệu z . Với ký hiệu, các công thức trở thành:

Bảng 2.34

Mẫu	$x = \bar{x} + zS$	$z = \frac{x - \bar{x}}{S}$
Quần thể	$x = \mu + z\sigma$	$z = \frac{x - \mu}{\sigma}$

Ví dụ 2.32**Vấn đề**

Hai học sinh, John và Ali, từ các trường trung học khác nhau, muốn tìm ai có GPA cao nhất khi so sánh với trường của mình. Học sinh nào có GPA cao nhất khi so sánh với trường của mình?

Bảng 2.35

Học sinh	GPA	GPA trung bình Trường	Độ lệch chuẩn Trường
John	2,85	3,0	0,7
Ali	77	80	10

Lời giải

Đối với mỗi học sinh, xác định có bao nhiêu độ lệch chuẩn (#ofSTDEVs) GPA của anh ấy cách xa trung bình, cho trường của anh ấy. Chú ý cẩn thận đến dấu khi so sánh và giải thích câu trả lời.

$$z = \#ofSTDEVs = \frac{\text{giá trị} - \text{trung bình}}{\text{độ lệch chuẩn}} = \frac{x - \mu}{\sigma}$$

$$\text{Đối với John, } z = \#ofSTDEVs = \frac{2,85 - 3,0}{0,7} = -0,21$$

$$\text{Đối với Ali, } z = \#ofSTDEVs = \frac{77 - 80}{10} = -0,3$$

John có GPA tốt hơn khi so sánh với trường của mình vì GPA của anh ấy **thấp hơn** 0,21 độ lệch chuẩn so với số bình quân trường của mình trong khi GPA của Ali **thấp hơn** 0,3 độ lệch chuẩn so với số bình quân trường của mình.

Điểm z (z-score) của John là -0,21 cao hơn điểm z (z-score) của Ali là -0,3. Đối với GPA, giá trị cao hơn tốt hơn, vậy chúng ta kết luận rằng John có GPA tốt hơn khi so sánh với trường của mình.

Bài tập vận dụng 2.32

Hai vận động viên bơi lội, Angie và Beth, từ các đội khác nhau, muốn tìm hiểu ai có thời gian nhanh nhất cho 50 mét tự do khi so sánh với đội của cô ấy. Vận động viên bơi nào có thời gian nhanh nhất khi so sánh với đội của cô ấy

Bảng 2.36

Vận động viên bơi	Thời gian (giây)	Thời gian Trung bình Đội	Độ lệch chuẩn Đội
Angie	26,2	27,2	0,8
Beth	27,3	30,1	1,4

Danh sách sau đây đưa ra một số thông tin giúp chúng ta hiểu rõ hơn về ý nghĩa của độ lệch chuẩn đối với phân phối dữ liệu.

Đối với BẤT KỲ tập dữ liệu nào, bất kể phân phối dữ liệu như thế nào:

- Ít nhất 75% dữ liệu nằm trong phạm vi hai độ lệch chuẩn tính từ giá trị trung bình.
- Ít nhất 89% dữ liệu nằm trong phạm vi ba độ lệch chuẩn tính từ giá trị trung bình.
- Ít nhất 95% dữ liệu nằm trong phạm vi 4,5 độ lệch chuẩn tính từ giá trị trung bình.
- Điều này được biết như là Định lý Chebyshev.

Đối với dữ liệu có Phân phối chuẩn, mà chúng ta sẽ nghiên cứu chi tiết sau này:

- Khoảng 68% dữ liệu nằm trong phạm vi một độ lệch chuẩn tính từ giá trị trung bình.
- Khoảng 95% dữ liệu nằm trong phạm vi hai độ lệch chuẩn tính từ giá trị trung bình.
- Hơn 99% dữ liệu nằm trong phạm vi ba độ lệch chuẩn tính từ giá trị trung bình.
- Điều này được gọi là Quy tắc thực nghiệm.
- Cần lưu ý rằng quy tắc này chỉ áp dụng khi hình dạng phân phối dữ liệu có dạng chuông và đối xứng. Chúng ta sẽ tìm hiểu thêm về điều này khi nghiên cứu phân phối xác suất "Chuẩn" hay "Gauss" trong các chương sau.

Hệ số biến Thiên

Một cách hữu ích khác để so sánh các phân phối, ngoài việc so sánh đơn giản các giá trị trung bình hoặc độ lệch chuẩn, là điều chỉnh theo sự khác biệt về thang đo của dữ liệu được đo lường. Đơn giản mà nói, một biến thiên lớn trong dữ liệu có giá trị trung bình lớn thì khác với cùng một biến thiên đó trong dữ liệu có giá trị trung bình nhỏ. Để điều chỉnh theo thang đo của dữ liệu cơ bản, Hệ số biến thiên (Coefficient of Variation - CV) đã được phát triển. Về mặt toán học:

$$CV = \frac{s}{\bar{x}} * 100$$
 với điều kiện $\bar{x} \neq 0$, trong đó s là độ lệch chuẩn của dữ liệu và $\bar{x}$ là giá trị trung bình.

Chúng ta có thể thấy rằng thước đo này đo lường mức độ biến thiên của dữ liệu cơ bản dưới dạng phần trăm của giá trị trung bình; tức là trọng tâm của tập dữ liệu. Thước đo này hữu ích trong việc so sánh rủi ro khi cần có sự điều chỉnh do sự khác biệt về thang đo của hai tập dữ liệu. Trên thực tế, thang đo được chuyển đổi thành một thang đo chung là tỷ lệ phần trăm, và cho phép so sánh trực tiếp mức độ biến thiên của hai hay nhiều tập dữ liệu khác nhau.

Thuật ngữ chính

Bảng tần suất (Frequency Table) một cách biểu diễn dữ liệu trong đó dữ liệu được nhóm để hiển thị cùng với các tần suất tương ứng.

Đa giác tần suất (Frequency Polygon) sử dụng đồ thị đường để biểu diễn tần suất của dữ liệu. Trục x là các điểm biên, trục y là tần suất (tương đối) tại các điểm biên.

Điểm giữa (Midpoint) trung bình của một khoảng trong bảng tần suất.

Đồ thị histogram (Histogram) một biểu diễn đồ thị thanh dưới dạng x - y của phân phối dữ liệu trong một tập dữ liệu; x đại diện cho dữ liệu và y đại diện cho tần suất, hoặc tần suất tương đối. Đồ thị bao gồm các hình chữ nhật liền kề.

Đồ thị hộp (box plot) một đồ thị cung cấp hình ảnh nhanh về 50% giữa của dữ liệu.

Độ lệch chuẩn (Standard Deviation) một số bằng căn bậc hai của phương sai và đo mức độ xa của các giá trị dữ liệu so với trung bình của chúng; ký hiệu: s cho độ lệch chuẩn mẫu và σ cho độ lệch chuẩn quần thể.

Giá trị ngoại vi (Outlier) một quan sát không phù hợp với phần còn lại của dữ liệu.

Khoảng tứ phân vị (Interquartile Range) hay *IQR*, là khoảng của 50 phần trăm giữa của các giá trị dữ liệu; *IQR* được tìm bằng cách trừ tứ phân vị thứ nhất khỏi tứ phân vị thứ ba.

Khoảng (Interval) còn được gọi là lớp. Một khoảng đại diện cho một phạm vi dữ liệu và được sử dụng khi hiển thị các tập dữ liệu lớn.

Lệch (Skewed) được sử dụng để mô tả dữ liệu không đối xứng; khi phía bên phải của biểu đồ trông "bị cắt bớt" so với phía bên trái, chúng ta nói nó "lệch trái". Khi phía bên trái của biểu đồ trông "bị cắt bớt" so với phía bên phải, chúng ta nói dữ liệu "lệch phải". Cách khác: khi các giá trị thấp hơn của dữ liệu phân tán rộng hơn, chúng ta nói dữ liệu lệch trái. Khi các giá trị lớn hơn phân tán rộng hơn, dữ liệu lệch phải.

Phân vị (Percentile) một số chia dữ liệu đã sắp xếp thành trăm phần; các phân vị có thể là hoặc không là một phần của dữ liệu. Trung vị của dữ liệu là tứ phân vị thứ hai hay phân vị thứ 50. Tứ phân vị thứ nhất và thứ ba lần lượt là phân vị thứ 25 và thứ 75.

Phương sai (Variance) trung bình của các bình phương độ lệch từ trung bình, hoặc bình phương của độ lệch chuẩn; đối với một tập dữ liệu, một độ lệch có thể được biểu diễn là $x - \bar{x}$ trong đó x là một giá trị của dữ liệu và $\bar{x}$ là trung bình mẫu. Phương sai mẫu bằng tổng bình phương các độ lệch chia cho hiệu của kích thước mẫu và một.

Tần suất tương đối (Relative Frequency) tỷ số của số lần một giá trị dữ liệu xuất hiện trong tập tất cả các kết quả so với số lượng tất cả các kết quả

Tần suất (Frequency) số lần một giá trị của dữ liệu xuất hiện

Trung bình cộng/ Bình quân cộng (Arithmetic Mean) một số đo sự hướng tâm của dữ liệu; tên gọi thông thường của trung bình là "trung bình cộng". Thuật ngữ "trung bình" là dạng rút gọn của "trung bình số học. Theo định nghĩa, trung bình của một mẫu (ký hiệu

bằng $\bar{x}$) $\bar{x} = \frac{\text{Tổng tất cả giá trị trong mẫu}}{\text{Số lượng giá trị trong mẫu}}$ và trung bình của một tổng thể (ký hiệu bằng μ) $\mu = \frac{\text{Tổng tất cả giá trị trong tổng thể}}{\text{Số lượng giá trị trong tổng thể}}$.

Trung bình nhân/ Bình quân nhân (Geometric Mean) là một số đo sự hướng tâm của dữ liệu, cho biết mức tăng trưởng trung bình cấp số nhân qua các giai đoạn thời gian.

Trung vị (Median) một số chia dữ liệu đã sắp xếp thành hai nửa; một nửa các giá trị bằng hoặc nhỏ hơn trung vị và một nửa các giá trị bằng hoặc lớn hơn trung vị. Trung vị có thể là hoặc không là một phần của dữ liệu.

Tứ phân vị thứ nhất (First Quartile) giá trị là trung vị của nửa dưới của tập dữ liệu đã sắp xếp

Tứ phân vị (Quartiles) các số chia dữ liệu thành bốn phần; các tứ phân vị có thể là hoặc không là một phần của dữ liệu. Tứ phân vị thứ hai là trung vị của dữ liệu.

Yếu vị (Mode) giá trị xuất hiện thường xuyên nhất trong một tập dữ liệu

Tóm tắt chương

2.1 Trình bày dữ liệu

Đồ thị thân và lá là một cách để vẽ biểu đồ dữ liệu và quan sát phân phối dữ liệu. Trong đồ thị thân và lá, tất cả các giá trị dữ liệu trong một lớp đều có thể nhìn thấy được. Ưu điểm của đồ thị thân và lá là tất cả các giá trị đều được liệt kê, khác với đồ thị histogram chỉ hiển thị các lớp dữ liệu. **Đồ thị đường** thường được sử dụng để biểu diễn một tập hợp các giá trị dữ liệu trong đó một đại lượng biến thiên theo thời gian. Những đồ thị này hữu ích để tìm ra xu hướng. Tức là, tìm ra một mẫu hình chung trong các tập dữ liệu, bao gồm nhiệt độ, doanh số, việc làm, lợi nhuận hoặc chi phí của công ty trong một khoảng thời gian. **Đồ thị thanh** là một biểu đồ sử dụng các thanh ngang hoặc dọc để so sánh giữa các phân loại. Một trục của biểu đồ hiển thị cụ thể các phân loại cần được so sánh, và trục kia biểu diễn một giá trị rời rạc. Một số đồ thị thanh trình bày các thanh được nhóm thành từng cụm nhiều hơn một (đồ thị thanh nhóm), và một số khác hiển thị các thanh được chia thành các phần để thể hiện tác động tích lũy (đồ thị thanh xếp chồng). Đồ thị thanh đặc biệt hữu ích khi sử dụng cho dữ liệu phân loại.

Đồ thị histogram là phiên bản đồ họa của bảng phân phối tần suất. Đồ thị bao gồm các thanh có chiều rộng bằng nhau được vẽ liền kề nhau. Trục hoành biểu diễn các lớp của giá trị dữ liệu định lượng và trục tung biểu diễn tần suất. Chiều cao của các thanh tương ứng với giá trị tần suất. Đồ thị histogram thường được sử dụng cho các tập lớn các dữ liệu định lượng, liên tục. Đa giác tần suất cũng có thể được sử dụng khi vẽ đồ thị các tập dữ liệu lớn với các điểm dữ liệu lặp lại. Dữ liệu thường được đặt trên trục y với tần suất được vẽ trên trục x. Đồ thị chuỗi thời gian có thể hữu ích khi quan sát lượng lớn dữ liệu cho một biến trong một khoảng thời gian.

2.2 Các số đo vị trí của dữ liệu

Các giá trị chia một tập dữ liệu đã sắp xếp theo thứ hạng thành 100 phần bằng nhau được gọi là **phân vị**. Phân vị được sử dụng để so sánh và diễn giải dữ liệu. Ví dụ, một quan sát

ở phân vị thứ 50 sẽ lớn hơn 50% các quan sát khác trong tập dữ liệu. Tứ phân vị chia dữ liệu thành các phần tư. Tứ phân vị thứ nhất (Q1) là phân vị thứ 25, tứ phân vị thứ hai (Q2 hoặc trung vị) là phân vị thứ 50, và tứ phân vị thứ ba (Q3) là phân vị thứ 75. Khoảng tứ phân vị, hay IQR, là khoảng của 50% giữa các giá trị dữ liệu. IQR được tìm bằng cách lấy Q3 trừ Q1, và có thể giúp xác định các giá trị ngoại vi bằng cách sử dụng hai biểu thức sau đây.

- $Q3 + IQR(1,5)$
- $Q1 - IQR(1,5)$

2.3 Các số đo tâm của dữ liệu

Số **bình quân** và số **trung vị** có thể được tính toán để giúp bạn tìm ra "tâm" của một tập dữ liệu. Số bình quân là ước lượng tốt nhất cho tập dữ liệu thực tế, nhưng số trung vị là thước đo tốt nhất khi một tập dữ liệu chứa nhiều giá trị ngoại vi hoặc giá trị cực đoan. Yếu vị sẽ cho bạn biết dữ liệu (hoặc các dữ liệu) xuất hiện thường xuyên nhất trong tập dữ liệu của bạn. Số bình quân, số trung vị và yếu vị cực kỳ hữu ích khi bạn cần phân tích dữ liệu của mình, nhưng nếu tập dữ liệu của bạn chỉ bao gồm các khoảng mà thiếu giá trị cụ thể, việc tính số bình quân có thể có vẻ không thể thực hiện được. Tuy nhiên, số bình quân có thể được xấp xỉ nếu bạn cộng biên dưới với biên trên và chia cho hai để tìm điểm giữa của mỗi khoảng. Nhân mỗi điểm giữa với số điểm dữ liệu được tìm thấy trong khoảng tương ứng. Chia tổng các giá trị này cho tổng số điểm dữ liệu trong tập hợp.

2.6 Độ lệch và bình quân, trung vị, và yếu vị

Quan sát phân phối của dữ liệu có thể tiết lộ nhiều điều về mối quan hệ giữa số bình quân, số trung vị và yếu vị. Có ba loại phân phối. Phân phối **lệch trái (hoặc âm)** có hình dạng như Hình 2.12. Phân phối **lệch phải (hoặc dương)** có hình dạng như Hình 2.13. Phân phối **đối xứng** trông như Hình 2.11.

2.7 Các số đo độ phân tán của dữ liệu

Độ lệch chuẩn có thể giúp bạn tính toán độ phân tán của dữ liệu. Có các công thức khác nhau để tính độ lệch chuẩn của một mẫu hoặc của một tổng thể.

- Độ lệch chuẩn cho phép chúng ta so sánh dữ liệu cá nhân hoặc các lớp với số bình quân của tập dữ liệu một cách bằng số.
- $s = \sqrt{\frac{\Sigma(x-\bar{x})^2}{n-1}}$ hoặc $s = \sqrt{\frac{\Sigma f(x-\bar{x})^2}{n-1}}$ là công thức để tính độ lệch chuẩn của một mẫu. Để tính độ lệch chuẩn của một tổng thể, chúng ta sẽ sử dụng số bình quân tổng thể, μ , và công thức $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$ hoặc $\sigma = \sqrt{\frac{\Sigma f(x-\mu)^2}{N}}$.

Tóm tắt công thức

2.2 Các số đo vị trí của Dữ liệu

$$i = \left(\frac{k}{100} \right) (n + 1)$$

trong đó i = thứ hạng hoặc vị trí của một giá trị dữ liệu,

k = phân vị thứ k ,

n = tổng số dữ liệu.

Biểu thức để tìm phân vị của một giá trị dữ liệu: $\left(\frac{x + 0,5y}{n} \right) (100)$

trong đó x = số điểm dữ liệu đếm từ dưới danh sách dữ liệu lên nhưng không bao gồm giá trị dữ liệu mà bạn muốn tìm phân vị,

y = số điểm dữ liệu bằng giá trị dữ liệu mà bạn muốn tìm phân vị,

n = tổng số dữ liệu

2.3 Các số đo tâm của dữ liệu

$\mu = \frac{\sum fm}{\sum f}$, trong đó f = tần suất khoảng và m = điểm giữa khoảng.

Số trung bình cộng cho một mẫu (ký hiệu bằng $\bar{x}$) là $\bar{x} = \frac{\text{Tổng tất cả giá trị trong mẫu}}{\text{Số lượng giá trị trong mẫu}}$

Số trung bình cộng cho một tổng thể (ký hiệu bằng μ) là $\mu = \frac{\text{Tổng tất cả giá trị trong quần thể}}{\text{Số lượng giá trị trong quần thể}}$

2.5 Trung bình nhân

Trung bình nhân:

$$\tilde{x} = \left(\prod_{i=1}^n x_i \right)^{\frac{1}{n}} = \sqrt[n]{x_1 \cdot x_2 \dots x_n} = (x_1 \cdot x_2 \dots x_n)^{\frac{1}{n}}$$

2.6 Độ lệch và bình quân, trung vị, và yếu vị

Công thức độ lệch: $a_3 = \frac{\sum (x_i - \bar{x})^3}{ns^3}$

Công thức Hệ số biến thiên: $CV = \frac{s}{\bar{x}} * 100$ với điều kiện $\bar{x} \neq 0$

2.7 Các số đo độ phân tán của dữ liệu

$$s_x = \sqrt{\frac{\sum f m^2}{n} - \bar{x}^2} \text{ trong đó}$$

s_x = độ lệch chuẩn mẫu

$\bar{x}$ = trung bình mẫu

$$\text{Công thức Độ lệch chuẩn mẫu } s = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} \text{ hoặc } s = \sqrt{\frac{\sum f(x - \bar{x})^2}{n-1}} \text{ hoặc } s = \sqrt{\frac{(\sum_{i=1}^n x^2) - n\bar{x}^2}{n-1}}$$

Đối với độ lệch chuẩn mẫu, mẫu số là $n - 1$, đó là kích thước mẫu trừ 1.

$$\text{Công thức Độ lệch chuẩn quần thể } \sigma = \sqrt{\frac{\sum (x - \mu)^2}{N}} \text{ hoặc } \sigma = \sqrt{\frac{\sum f(x - \mu)^2}{N}} \text{ hoặc } \sigma = \sqrt{\frac{\sum_{i=1}^N x_i^2}{N} - \mu^2}$$

Đối với độ lệch chuẩn quần thể, mẫu số là N , tức là số lượng phần tử trong quần thể.

Tài liệu tham khảo

2.1 Trình bày dữ liệu

Doyle, Brandon. "TikTok Statistics – Updated Mar 2023." Wallaroo. March 21, 2023. <https://wallaroomedia.com/blog/social-media/tiktok-statistics/>.

Table 219.20. Public High School Graduates, by Region, State, and Jurisdiction: Selected Years, 1980-81 through 2030-31."

National Center for Education Statistics. Accessed May 24, 2023. https://nces.ed.gov/programs/digest/d21/tables/dt21_219.20.asp.

"Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at <http://www.cdc.gov/obesity/data/adult.html> (accessed September 13, 2013).

Data on annual homicides in Detroit, 1961–73, from Gunst & Mason's book 'Regression Analysis and its Application', Marcel Dekker

"Timeline: Guide to the U.S. Presidents: Information on every president's birthplace, political party, term of office, and more." Scholastic, 2013. Available online at <http://www.scholastic.com/teachers/article/timeline-guide-us-presidents> (accessed April 3, 2013).

"Presidents." Fact Monster. Pearson Education, 2007. Available online at <http://www.factmonster.com/ipka/A0194030.html> (accessed April 3, 2013).

"Food Security Statistics." Food and Agriculture Organization of the United Nations. Available online at <http://www.fao.org/economic/ess/ess-fs/en/> (accessed April 3, 2013).

"Consumer Price Index." United States Department of Labor: Bureau of Labor Statistics. Available online at <http://data.bls.gov/pdq/SurveyOutputServlet> (accessed April 3, 2013).

"CO2 emissions (kt)." The World Bank, 2013. Available online at <http://databank.worldbank.org/data/home.aspx> (accessed April 3, 2013).

"Births Time Series Data." General Register Office For Scotland, 2013. Available online at <http://www.gro-scotland.gov.uk/statistics/theme/vital-events/births/time-series.html> (accessed April 3, 2013).

"Demographics: Children under the age of 5 years underweight." Indexmundi. Available online at <http://www.indexmundi.com/g/r.aspx?t=50&v=2224&aml=en> (accessed April 3, 2013).

Gunst, Richard, Robert Mason. Regression Analysis and Its Application: A Data-Oriented Approach. CRC Press: 1980.

"Overweight and Obesity: Adult Obesity Facts." Centers for Disease Control and Prevention. Available online at <http://www.cdc.gov/obesity/data/adult.html> (accessed September 13, 2013).

2.2 Các số đo vị trí của dữ liệu

Cauchon, Dennis, Paul Overberg. "Census data shows minorities now a majority of U.S. births." USA Today, 2012. Available online at <http://usatoday30.usatoday.com/news/nation/story/2012-05-17/minority-birthscensus/55029100/1> (accessed April 3, 2013).

Data from the United States Department of Commerce: United States Census Bureau. Available online at <http://www.census.gov/> (accessed April 3, 2013).

"1990 Census." United States Department of Commerce: United States Census Bureau. Available online at <http://www.census.gov/main/www/cen1990.html> (accessed April 3, 2013).

Data from San Jose Mercury News.

Data from Time Magazine; survey by Yankelovich Partners, Inc.

2.3 Các số đo tâm của dữ liệu

Data from The World Bank, available online at <http://www.worldbank.org> (accessed April 3, 2013).

"Demographics: Obesity – adult prevalence rate." Indexmundi. Available online at <http://www.indexmundi.com/g/r.aspx?t=50&v=2228&l=en> (accessed April 3, 2013).

2.7 Các số đo độ phân tán của dữ liệu

Data from Microsoft Bookshelf.

King, Bill. "Graphically Speaking." Institutional Research, Lake Tahoe Community College. Available online at <http://www.ltcc.edu/web/about/institutional-research> (accessed April 3, 2013).

3 XÁC SUẤT



Hình 3.1 Mưa thiên thạch là hiện tượng rất hiếm, nhưng xác suất xảy ra của chúng có thể tính được. (Nguồn: chỉnh sửa từ tác phẩm "Một thiên thạch trong đỉnh cao của Trận mưa thiên thạch Leonid năm 2009" bởi Navicore/ Wikimedia Commons, CC BY 3.0)

Giới thiệu

Chúng ta thường xuyên cần thiết phải "đoán" về kết quả của một sự việc để đưa ra quyết định. Các chính trị gia nghiên cứu các cuộc thăm dò ý kiến để đoán khả năng thắng cử của họ. Giáo viên lựa chọn nội dung cụ thể cho môn học dựa trên những gì họ nghĩ học sinh có thể hiểu được. Bác sĩ chọn các phương pháp điều trị cần thiết cho các bệnh khác nhau dựa trên đánh giá của họ về kết quả có thể xảy ra. Một số người thích chơi cá cược là vì tin rằng khả năng thắng là tốt. Bạn cũng có thể đã chọn ngành học của mình dựa trên khả năng có việc làm trong tương lai.

Rất có thể bạn đã từng sử dụng xác suất. Thực vậy, có xác suất rằng bạn có cảm nhận trực quan về xác suất. Xác suất liên quan đến cơ hội xảy ra của một sự việc. Bất cứ khi nào bạn cân nhắc khả năng có nên làm bài tập về nhà hay học cho một kỳ thi, bạn đang sử dụng xác suất. Trong chương này, bạn sẽ học cách giải quyết các bài toán xác suất sử dụng một cách tiếp cận có hệ thống.

3.1 Thuật ngữ

Xác suất là một số đo về mức độ chắc chắn của chúng ta về kết quả của một **thực nghiệm**⁴ (experiment) hoặc hoạt động cụ thể. Một thực nghiệm là một hoạt động có kế hoạch được thực hiện trong điều kiện có kiểm soát. Nếu kết quả không được xác định trước, thì thí

⁴ còn gọi là "thí nghiệm"

thực nghiệm đó được gọi là **thực nghiệm cơ hội**⁵ (chance experiment). Tung một đồng xu hai lần là ví dụ về một thực nghiệm.

Kết quả của một thực nghiệm được gọi là một **phương án**⁶ (outcome). **Không gian mẫu** (sample space) của một thực nghiệm là tập hợp tất cả các phương án có thể có. Có ba cách để biểu diễn không gian mẫu là: liệt kê các phương án có thể có, tạo sơ đồ cây, hoặc tạo sơ đồ Venn. Chữ in hoa S được sử dụng để ký hiệu không gian mẫu. Ví dụ, nếu bạn tung một đồng xu, $S = \{H, T\}$ trong đó H = mặt ngửa và T = mặt úp là các phương án.

Một **sự kiện**⁷ (event) là bất kỳ sự kết hợp nào của các phương án. Các chữ cái in hoa như A và B biểu diễn các sự kiện. Ví dụ, nếu thực nghiệm là tung một đồng xu, sự kiện A có thể là nhận được nhiều nhất một mặt ngửa. Xác suất của một sự kiện A được viết là $P(A)$.

Xác suất của bất kỳ phương án nào là **tần suất tương đối trong dài hạn** (long-term relative frequency) của phương án đó. **Xác suất nằm trong khoảng từ 0 đến 1, bao gồm cả 0 và 1** (nghĩa là, 0 và 1 và tất cả các số giữa các giá trị này). $P(A) = 0$ có nghĩa là sự kiện A không bao giờ xảy ra. $P(A) = 1$ có nghĩa là sự kiện A luôn xảy ra. $P(A) = 0,5$ có nghĩa là sự kiện A có khả năng xảy ra hoặc không xảy ra như nhau. Ví dụ, nếu bạn tung một đồng xu công bằng nhiều lần (từ 20 đến 2.000 đến 20.000 lần) thì tần suất tương đối của mặt ngửa tiến gần đến 0,5 (xác suất của mặt ngửa).

Đồng khả năng (equally likely) có nghĩa là mỗi phương án của một thực nghiệm xảy ra với xác suất bằng nhau. Ví dụ, nếu bạn tung một con xúc xắc công bằng sáu mặt, mỗi mặt (1, 2, 3, 4, 5, hoặc 6) đều có khả năng xuất hiện như bất kỳ mặt nào khác. Nếu bạn tung một đồng xu công bằng, mặt ngửa (H) và mặt úp (T) đều có khả năng xuất hiện như nhau. Nếu bạn đoán ngẫu nhiên câu trả lời cho một câu hỏi đúng/sai trong bài kiểm tra, bạn có khả năng chọn câu trả lời đúng hoặc câu trả lời sai như nhau.

Để tính xác suất của một sự kiện A khi tất cả các phương án trong không gian mẫu đều đồng khả năng, hãy đếm số phương án thuộc sự kiện A và chia cho tổng số phương án trong không gian mẫu. Ví dụ, nếu bạn tung một đồng xu 10 xu và một đồng xu 5 xu công bằng, không gian mẫu là $\{HH, TH, HT, TT\}$ trong đó T = mặt úp và H = mặt ngửa. Không gian mẫu có bốn phương án. Nếu sự kiện A = nhận được một mặt ngửa. Có hai phương án thỏa mãn điều kiện này $\{HT, TH\}$, vì vậy $P(A) = \frac{2}{4} = 0,5$.

Giả sử bạn tung một con xúc xắc sáu mặt công bằng, với các số $\{1, 2, 3, 4, 5, 6\}$ trên các mặt của nó. Gọi sự kiện E = tung ra mặt không nhỏ hơn 5. Có hai phương án $\{5, 6\}$. $P(E) = \frac{2}{6}$. Nếu bạn chỉ tung con xúc xắc một vài lần, bạn không nên ngạc nhiên nếu kết quả quan sát được không khớp với xác suất. Nếu bạn tung con xúc xắc rất nhiều lần, bạn sẽ thấy rằng, nhìn chung, $\frac{2}{6}$ số lần tung sẽ cho kết quả "ít nhất là năm". Bạn không nên mong đợi là chính xác $\frac{2}{6}$. Tần suất tương đối dài hạn của việc thu được kết quả này sẽ tiến gần đến xác suất lý thuyết $\frac{2}{6}$ khi số lần lặp lại tăng lên.

⁵ Còn gọi là "thực nghiệm may rủi"

⁶ Còn gọi là "kết quả".

⁷ Còn gọi là "biến cố"

Đặc tính quan trọng trên của các thực nghiệm xác suất được gọi là **Luật số lớn** (law of large numbers), nó nêu rằng khi số lần lặp lại một thực nghiệm tăng lên, tần suất tương đối thu được từ thực nghiệm có xu hướng ngày càng tiến gần đến xác suất lý thuyết. Mặc dù các kết quả không xảy ra theo bất kỳ mẫu hoặc thứ tự cố định nào, nhưng nhìn chung, tần suất tương đối quan sát được trong dài hạn sẽ tiến gần đến xác suất lý thuyết. (Thuật ngữ **thực nghiệm** thường được sử dụng thay cho thuật ngữ quan sát).

Rất cần phải lưu ý rằng là trong nhiều trường hợp, các phương án của thực nghiệm không đồng khả năng. Một đồng xu hoặc xúc xắc có thể không công bằng, hay **thiên lệch** (bias). Hai giáo sư toán ở châu Âu đã yêu cầu sinh viên thống kê của họ kiểm tra đồng xu một Euro của Bỉ và phát hiện ra rằng trong 250 lần thử, mặt ngửa xuất hiện 56% thời gian và mặt úp xuất hiện 44% thời gian. Dữ liệu này dường như cho thấy đồng xu không phải là đồng xu công bằng; lặp lại nhiều lần sẽ giúp đưa ra kết luận chính xác hơn về sự thiên lệch đó. Một số xúc xắc có thể bị thiên lệch. Hãy nhìn vào xúc xắc trong trò chơi bạn có ở nhà; các chấm trên mỗi mặt thường là các lỗ nhỏ được khắc ra và sau đó được sơn để làm cho các chấm hiển thị rõ ràng. Xúc xắc của bạn có thể bị thiên lệch hoặc không; có khả năng các kết quả có thể bị ảnh hưởng bởi sự khác biệt trọng lượng nhỏ do số lượng lỗ khác nhau trên các mặt. Các sòng bạc kiếm được tiền phụ thuộc vào kết quả từ việc tung xúc xắc, vì vậy xúc xắc sòng bạc được làm theo cách khác để loại bỏ sự thiên lệch. Xúc xắc sòng bạc có các mặt phẳng; các lỗ được lấp đầy hoàn toàn bằng sơn có cùng khối lượng riêng với vật liệu làm xúc xắc để mỗi mặt có khả năng xuất hiện như nhau. Phần sau chúng ta sẽ học các kỹ thuật xử lý xác suất của các sự kiện không đồng khả năng.

Sự kiện "HOẶC" (OR event):

Một phương án thuộc sự kiện A HOẶC B nếu phương án đó thuộc A hoặc thuộc B hoặc thuộc cả A và B.

Sự kiện A HOẶC B còn có thể được viết là A Hợp B, với ký hiệu như sau:

$$A \cup B$$

Ví dụ, cho $A = \{1, 2, 3, 4, 5\}$ và $B = \{4, 5, 6, 7, 8\}$.

Khi đó $A \cup B = \{1, 2, 3, 4, 5, 6, 7, 8\}$. Lưu ý rằng 4 và 5 KHÔNG được liệt kê hai lần.

Sự kiện "VÀ" (AND event):

Một phương án thuộc sự kiện A VÀ B nếu phương án đó thuộc cả A và B cùng một lúc.

Sự kiện A VÀ B còn có thể được viết là A Giao B, với ký hiệu như sau

$$A \cap B$$

Ví dụ, cho A và B lần lượt là $\{1, 2, 3, 4, 5\}$ và $\{4, 5, 6, 7, 8\}$.

Khi đó $A \cap B = \{4, 5\}$.

Phần bù (complement) của sự kiện A được ký hiệu là A' . A' bao gồm tất cả các phương án KHÔNG thuộc A. Lưu ý rằng $P(A) + P(A') = 1$. Ví dụ, cho $S = \{1, 2, 3, 4, 5, 6\}$ và cho $A = \{1, 2, 3, 4\}$. Khi đó, $A' = \{5, 6\}$. $P(A) = \frac{4}{6}$, $P(A') = \frac{2}{6}$, và $P(A) + P(A') = \frac{4}{6} + \frac{2}{6} = 1$

Xác suất có điều kiện của A với điều kiện B được viết là $P(A|B)$. $P(A|B)$ là xác suất sự kiện A sẽ xảy ra với điều kiện sự kiện B đã xảy ra. **Một điều kiện thu sẽ hẹp không gian mẫu.** Chúng ta tính xác suất của A từ không gian mẫu thu hẹp B. Công thức để tính $P(A|B)$ is $P(A|B) = \frac{P(A \text{ AND } B)}{P(B)}$ trong đó $P(B)$ lớn hơn không.

Ví dụ, giả sử chúng ta tung một con xúc xắc sáu mặt công bằng. Không gian mẫu $S = \{1, 2, 3, 4, 5, 6\}$. Cho A = sự kiện tung ra mặt là 2 hoặc 3 và B = sự kiện tung ra mặt là số chẵn (2, 4, 6). Để tính $P(A|B)$, chúng ta đếm số phương án 2 hoặc 3 trong không gian mẫu $B = \{2, 4, 6\}$. Sau đó, chúng ta chia số đó cho số phương án của B (thay vì S).

Chúng ta nhận được cùng kết quả bằng cách sử dụng công thức. Hãy nhớ rằng S có sáu phương án.

$$P(A|B) = \frac{P(A \text{ AND } B)}{P(B)} = \frac{\frac{(\text{số phương án là 2 hoặc 3 và chẵn trong } S)}{6}}{\frac{(\text{số phương án là chẵn trong } S)}{6}} = \frac{\frac{2}{6}}{\frac{3}{6}} = \frac{2}{3}$$

Tỷ lệ cược

Tỷ lệ cược của một sự kiện biểu thị xác suất dưới dạng tỷ số giữa thành công và thất bại. Khái niệm này được sử dụng phổ biến trong nhiều ứng dụng thực tế. Về mặt toán học, tỷ lệ cược của một sự kiện được định nghĩa là:

$$\frac{P(A)}{1 - P(A)}$$

trong đó $P(A)$ là xác suất thành công và $1 - P(A)$ là xác suất thất bại. Tỷ số cược thường được biểu diễn dưới dạng "tỷ số : mẫu số", chẳng hạn 2:1. Trong trường hợp này, xác suất thắng gấp đôi xác suất thua, do đó xác suất thắng là 0,66. Một xác suất thắng là 0,60 sẽ cho tỷ lệ cược ủng hộ việc thắng là 3:2. Mặc dù việc tính toán tỷ lệ cược có ứng dụng trong một số lĩnh vực, nhưng lại ít có giá trị trong việc hiểu lý thuyết xác suất và thống kê.

Hiểu các thuật ngữ và ký hiệu

Khi giải bài toán xác suất cần đọc kỹ bài toán để hiểu đâu là sự kiện. Hiểu rõ câu từ là là bước quan trọng đầu tiên. Đọc lại bài toán nhiều lần nếu cần thiết; xác định rõ đâu là sự kiện cần quan tâm. Xác định xem có điều kiện nào được nêu trong nội dung cho thấy xác suất là có điều kiện hay không; xác định cẩn thận điều kiện, nếu có.

Ví dụ 3.1

Vấn đề

Không gian mẫu S là các số nguyên bắt đầu từ 1 và nhỏ hơn 20.

a. $S =$ _____

Cho sự kiện A = các số chẵn và sự kiện B = các số lớn hơn 13.

b. $A =$ _____, $B =$ _____

c. $P(A) =$ _____, $P(B) =$ _____

- d. $A \text{ AND } B = \underline{\hspace{2cm}}$, $A \text{ OR } B = \underline{\hspace{2cm}}$
- e. $P(A \text{ AND } B) = \underline{\hspace{2cm}}$, $P(A \text{ OR } B) = \underline{\hspace{2cm}}$
- f. $A' = \underline{\hspace{2cm}}$, $P(A') = \underline{\hspace{2cm}}$
- g. $P(A) + P(A') = \underline{\hspace{2cm}}$
- h. $P(A|B) = \underline{\hspace{2cm}}$, $P(B|A) = \underline{\hspace{2cm}}$; các xác suất có bằng nhau không?

Lời giải

- a. $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19\}$
- b. $A = \{2, 4, 6, 8, 10, 12, 14, 16, 18\}$, $B = \{14, 15, 16, 17, 18, 19\}$
- c. $P(A) = \frac{9}{19}$, $P(B) = \frac{6}{19}$
- d. $A \text{ VÀ } B = \{14, 16, 18\}$, $A \text{ HOẶC } B = \{2, 4, 6, 8, 10, 12, 14, 15, 16, 17, 18, 19\}$
- e. $P(A \text{ VÀ } B) = \frac{3}{19}$, $P(A \text{ HOẶC } B) = \frac{12}{19}$
- f. $A' = 1, 3, 5, 7, 9, 11, 13, 15, 17, 19$; $P(A') = \frac{10}{19}$
- g. $P(A) + P(A') = 1$ ($\frac{9}{19} + \frac{10}{19} = 1$)
- h. $P(A|B) = \frac{P(A \text{ VÀ } B)}{P(B)} = \frac{3}{6}$, $P(B|A) = \frac{P(A \text{ VÀ } B)}{P(A)} = \frac{3}{9}$, Các xác suất không bằng nhau

Bài tập vận dụng 3.1

- a. Vấn đề
- b. Không gian mẫu S là tất cả các cặp có thứ tự của hai số nguyên, số thứ nhất từ 1 đến 3 và số thứ hai từ 1 đến 4 (Ví dụ: $(1, 4)$).
- c. $S = \underline{\hspace{2cm}}$
- Gọi sự kiện A = tổng là số chẵn và sự kiện B = số thứ nhất là số nguyên tố.
- d. $A = \underline{\hspace{2cm}}$, $B = \underline{\hspace{2cm}}$
- e. $P(A) = \underline{\hspace{2cm}}$, $P(B) = \underline{\hspace{2cm}}$
- f. $A \text{ AND } B = \underline{\hspace{2cm}}$, $A \text{ OR } B = \underline{\hspace{2cm}}$
- g. $P(A \text{ AND } B) = \underline{\hspace{2cm}}$, $P(A \text{ OR } B) = \underline{\hspace{2cm}}$
- h. $B' = \underline{\hspace{2cm}}$, $P(B') = \underline{\hspace{2cm}}$
- i. $P(A) + P(A') = \underline{\hspace{2cm}}$
- j. $P(A|B) = \underline{\hspace{2cm}}$, $P(B|A) = \underline{\hspace{2cm}}$; các xác suất có bằng nhau không?

Ví dụ 3.2

Vấn đề

Một con xúc xắc sáu mặt công bằng được tung. Mô tả không gian mẫu S , xác định mỗi sự kiện sau đây bằng một tập con của S và tính xác suất của nó (phương án là số chấm được tung ra).

- a. Sự kiện T = kết quả là 2.
- b. Sự kiện A = kết quả là một số chẵn.
- c. Sự kiện B = kết quả nhỏ hơn 4.
- d. Phần bù của A .
- e. $A \mid B$
- f. $B \mid A$
- g. $A \cap B$
- h. $A \cup B$
- i. $A \cup B'$
- j. Sự kiện N = kết quả là một số nguyên tố.
- k. Sự kiện I = kết quả là bảy.

Lời giải

- a. $T = \{2\}, P(T) = \frac{1}{6}$
- b. $A = \{2, 4, 6\}, P(A) = \frac{1}{2}$
- c. $B = \{1, 2, 3\}, P(B) = \frac{1}{2}$
- d. $A' = \{1, 3, 5\}, P(A') = \frac{1}{2}$
- e. $A|B = \{2\}, P(A|B) = \frac{1}{3}$
- f. $B|A = \{2\}, P(B|A) = \frac{1}{3}$
- g. $A \text{ VÀ } B = \{2\}, P(A \text{ VÀ } B) = \frac{1}{6}$
- h. $A \text{ HOẶC } B = \{1, 2, 3, 4, 6\}, P(A \text{ HOẶC } B) = \frac{5}{6}$
- i. $A \text{ HOẶC } B' = \{2, 4, 5, 6\}, P(A \text{ HOẶC } B') = \frac{2}{3}$
- j. $N = \{2, 3, 5\}, P(N) = \frac{1}{2}$
- k. Một con xúc xắc sáu mặt không có bảy chấm. $P(7) = 0$.

Bài tập vận dụng 3.2

Một số được chọn ngẫu nhiên trong khoảng từ 1 đến 10. Mô tả không gian mẫu S , xác định mỗi sự kiện sau đây bằng một tập con của S , và tính xác suất của nó (phương án là số được chọn).

- Sự kiện F = kết quả là 5.
- Sự kiện A = số lớn hơn 6.
- Sự kiện B = số là số lẻ.
- Phần bù của B .
- A VỚI ĐIỀU KIỆN B .
- B VỚI ĐIỀU KIỆN A .
- A HOẶC B .
- A' HOẶC B .
- Sự kiện P = kết quả là một số hợp.
- Sự kiện M = số là bội số của 3.

Ví dụ 3.3

Vấn đề

BẢNG 3.1 mô tả phân phối của một mẫu ngẫu nhiên S gồm 100 cá nhân, được tổ chức theo giới tính khi sinh và việc họ thuận tay phải hay tay trái.

Bảng 3.1

	Thuận tay phải	Thuận tay trái
Nam	43	9
Nữ	44	4

Đặt M = sự kiện đối tượng là nam, F = sự kiện đối tượng là nữ, R = sự kiện đối tượng thuận tay phải, L = đối tượng thuận tay trái. Tính các xác suất sau:

- $P(M)$
- $P(F)$
- $P(R)$
- $P(L)$
- $P(M \text{ AND } R)$
- $P(F \text{ AND } L)$
- $P(M \text{ OR } F)$
- $P(M \text{ OR } R)$

i. $P(F \text{ OR } L)$

j. $P(M')$

k. $P(R|M)$

l. $P(F|L)$

m. $P(L|F)$

Lời giải

a. $P(M) = 0,52$

b. $P(F) = 0,48$

c. $P(R) = 0,87$

d. $P(L) = 0,13$

e. $P(M \text{ AND } R) = 0,43$

f. $P(F \text{ AND } L) = 0,04$

g. $P(M \text{ OR } F) = 1$

h. $P(M \text{ OR } R) = 0,96$

i. $P(F \text{ OR } L) = 0,57$

j. $P(M') = 0,48$

k. $P(R|M) = 0,8269$ (làm tròn đến 4 chữ số thập phân)

l. $P(F|L) = 0,3077$ (làm tròn đến 4 chữ số thập phân)

m. $P(L|F) = 0,0833$

Bài tập vận dụng 3.3

Bảng này mô tả phân phối của một mẫu ngẫu nhiên S gồm 100 cá nhân, được tổ chức theo giới tính khi sinh và liệu họ thích trà hay cà phê.

Bảng 3.2

	Trà	Cà phê
Nam	22	26
Nữ	16	36

Đặt M = sự kiện đối tượng là nam, W = sự kiện đối tượng là nữ, T = sự kiện đối tượng thích trà, và C = sự kiện đối tượng thích cà phê. Tính các xác suất sau:

a. $P(M)$

b. $P(W)$

c. $P(T)$

d. $P(C)$

- e. $P(M \text{ VÀ } T)$
 - f. $P(W \text{ VÀ } C)$
 - g. $P(M \text{ HOẶC } W)$
 - h. $P(M \text{ HOẶC } T)$
 - i. $P(W \text{ HOẶC } C)$
 - j. $P(M')$
 - k. $P(T|M)$
 - l. $P(W|C)$
 - m. $P(C|W)$
-

3.2 Các sự kiện độc lập và xung khắc

Độc lập và xung khắc **không** có cùng ý nghĩa.

Các sự kiện độc lập (independent event)

Hai sự kiện được gọi là độc lập nếu những điều sau đây là đúng:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \text{ AND } B) = P(A)P(B)$

Hai sự kiện A và B là các sự kiện độc lập nếu việc biết một sự kiện xảy ra không ảnh hưởng đến xác suất sự kiện kia xảy ra. Ví dụ, kết quả của hai lần tung một con xúc xắc công bằng là các sự kiện độc lập. Kết quả của lần tung đầu tiên không làm thay đổi xác suất cho kết quả của lần tung thứ hai. Để chứng minh hai sự kiện là độc lập, bạn chỉ cần chứng minh **một** trong các điều kiện trên. Nếu hai sự kiện **KHÔNG** độc lập, thì chúng ta nói rằng chúng là các **sự kiện phụ thuộc** (dependent event).

Việc lấy mẫu có thể được thực hiện **có** hoàn lại hoặc **không** hoàn lại.

- **Có hoàn lại** (with replacement): Nếu mỗi phần tử của một quần thể được hoàn lại (thay thế) sau khi nó được chọn, thì phần tử đó có khả năng được chọn nhiều hơn một lần. Khi lấy mẫu được thực hiện có hoàn lại, thì các sự kiện được coi là độc lập, nghĩa là kết quả của lần chọn đầu tiên sẽ không thay đổi xác suất cho lần chọn thứ hai.
- **Không hoàn lại** (without replacement): Khi lấy mẫu được thực hiện không hoàn lại, mỗi phần tử của quần thể chỉ có thể được chọn một lần. Trong trường hợp này, xác suất cho lần chọn thứ hai bị ảnh hưởng bởi kết quả của lần chọn đầu tiên. Các sự kiện được coi là phụ thuộc hoặc không độc lập.

Nếu không biết liệu A và B có độc lập hay phụ thuộc, **giả định chúng là phụ thuộc cho đến khi bạn có thể chứng minh ngược lại.**

Ví dụ 3.4

Bạn có một bộ bài 52 lá được xáo trộn kỹ và công bằng. Nó bao gồm bốn chất. Các chất là chuồn (clubs), rô (diamonds), cơ (hearts) và bích (spades). Có 13 lá bài trong mỗi chất bao gồm 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (bồi), Q (đằm), K (già) của chất đó.

a. Lấy mẫu có hoàn lại:

Giả sử bạn rút ba lá bài có hoàn lại. Lá bài đầu tiên bạn rút ra từ 52 lá bài là Q bích. Bạn đặt lá bài này trở lại, xáo trộn lại bài và rút lá bài thứ hai từ bộ bài 52 lá. Đó là lá 10 chuồn. Bạn đặt lá bài này trở lại, xáo trộn lại bài và rút lá bài thứ ba từ bộ bài 52 lá. Lần này, lá bài là Q bích một lần nữa. Kết quả rút bài của bạn là {Q bích, 10 chuồn, Q bích}. Bạn đã rút được Q bích hai lần. Bạn rút mỗi lá bài từ bộ bài 52 lá.

b. Lấy mẫu không hoàn lại:

Giả sử bạn rút ba lá bài không hoàn lại. Lá bài đầu tiên bạn rút ra từ 52 lá bài là K cơ. Bạn để lá bài này sang một bên và rút lá bài thứ hai từ 51 lá bài còn lại trong bộ bài. Đó là lá 3 rô. Bạn để lá bài này sang một bên và rút lá bài thứ ba từ 50 lá bài còn lại trong bộ bài. Lá bài thứ ba là J bích. Kết quả rút bài của bạn là {K cơ, 3 rô, J bích}. Vì bạn đã rút bài không hoàn lại, bạn không thể rút cùng một lá bài hai lần.

Bài tập vận dụng 3.4

Bạn có một bộ bài 52 lá được xáo trộn kỹ và công bằng. Nó bao gồm bốn chất. Các chất là chuồn, rô, cơ và bích. Có 13 lá bài trong mỗi chất bao gồm 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (bồi), Q (đằm), K (già) của chất đó. Ba lá bài được rút ngẫu nhiên.

a. Giả sử bạn biết rằng các lá bài được rút là Q bích, K cơ và Q bích. Bạn có thể quyết định liệu việc lấy mẫu là có hoàn lại hay không hoàn lại không?

b. Giả sử bạn biết rằng các lá bài được rút là Q bích, K cơ và J bích. Bạn có thể quyết định liệu việc lấy mẫu là có hoàn lại hay không hoàn lại không? Giả sử bạn rút ba lá bài không hoàn lại. Lá bài đầu tiên bạn rút ra từ 52 lá bài là K cơ. Bạn để lá bài này sang một bên và rút lá bài thứ hai từ 51 lá bài còn lại trong bộ bài. Đó là lá 3 rô. Bạn để lá bài này sang một bên và rút lá bài thứ ba từ 50 lá bài còn lại trong bộ bài. Lá bài thứ ba là J bích. Kết quả rút bài của bạn là {K cơ, 3 rô, J bích}. Vì bạn đã rút bài không hoàn lại, bạn không thể rút cùng một lá bài hai lần.

Ví dụ 3.5

Vấn đề

Bạn có một bộ bài 52 lá được xáo trộn kỹ và công bằng. Nó bao gồm bốn chất. Các chất là chuồn, rô, cơ và bích. Có 13 lá bài trong mỗi chất bao gồm 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (bồi), Q (đằm) và K (già) của chất đó. S = bích, H = Cơ, D = Rô, C = Chuồn.

a. Giả sử bạn rút bốn lá bài, nhưng không đặt lá bài nào trở lại bộ bài. Các lá bài của bạn là QS, 1D, 1C, QD.

b. Giả sử bạn rút bốn lá bài và đặt mỗi lá bài trở lại trước khi bạn rút lá bài tiếp theo. Các lá bài của bạn là KH, 7D, 6D, KH.

Trong a. hay b., trường hợp nào bạn đã lấy mẫu có hoàn lại, và trường hợp nào bạn đã lấy mẫu không hoàn lại?

Lời giải

a. Không hoàn lại;

b. Có hoàn lại

Bài tập vận dụng 3.5

Bạn có một bộ bài 52 lá được xáo trộn kỹ và công bằng. Nó bao gồm 4 chất. Các chất là chuồn, rô, cơ và bích. Có 13 lá bài trong mỗi chất bao gồm 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, J (bồi), Q (đằm) và K (già) của chất đó. S = bích, H = Cơ, D = Rô, C = Chuồn. Giả sử bạn lấy mẫu bốn lá bài không hoàn lại. Kết quả nào sau đây là có thể? Trả lời cùng câu hỏi cho việc lấy mẫu có hoàn lại.

a. QS, 1D, 1C, QD

b. KH, 7D, 6D, KH

c. QS, 7D, 6D, KS

Các sự kiện xung khắc⁸ (mutually exclusive event)

A và B là các sự kiện xung khắc nếu chúng không thể xảy ra cùng một lúc. Điều này có nghĩa là A và B không có phương án chung nào và $P(A \text{ VÀ } B) = 0$.

Ví dụ, giả sử không gian mẫu $S = \{1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$. Cho $A = \{1, 2, 3, 4, 5\}$, $B = \{4, 5, 6, 7, 8\}$ và $C = \{7, 9\}$. $A \cap B = \{4, 5\}$. $P(A \cap B) = \frac{2}{10}$ và không bằng 0. Do đó, A và B không xung khắc. A và C không có số nào chung nên $P(A \cap C) = 0$. Do đó, A và C là xung khắc.

Nếu không biết liệu A và B có xung khắc hay không, **giả định chúng không xung khắc cho đến khi bạn có thể chứng minh ngược lại**. Các ví dụ sau minh họa các định nghĩa và thuật ngữ này.

Ví dụ 3.6

Tung hai đồng xu công bằng. (Đây là một thực nghiệm.)

Không gian mẫu là $\{HH, HT, TH, TT\}$ trong đó T = úp và H = ngửa. Các phương án là HH , HT , TH và TT . Các phương án HT và TH là khác nhau. HT có nghĩa là đồng xu thứ nhất

⁸ Còn gọi là “biến cố loại trừ lẫn nhau”

ra mặt ngửa và đồng xu thứ hai ra mặt úp. TH có nghĩa là đồng xu thứ nhất ra mặt úp và đồng xu thứ hai ra mặt ngửa.

- Gọi A = sự kiện nhận được nhiều nhất một mặt úp. (Nhiều nhất một mặt úp có nghĩa là không có hoặc một mặt úp từ 2 đồng xu.) Khi đó A có thể được viết là {HH, HT, TH}. Phương án HH cho thấy không có mặt úp. HT và TH mỗi phương án cho thấy một mặt úp.
- Gọi B = sự kiện nhận được tất cả mặt úp. B có thể được viết là {TT}. B là phần bù (complement) của A, nên $B = A'$. Ngoài ra, $P(A) + P(B) = P(A) + P(A') = 1$.
- Xác suất cho A và cho B là $P(A) = \frac{3}{4}$ và $P(B) = \frac{1}{4}$.
- Gọi C = sự kiện nhận được tất cả mặt ngửa. $C = \{HH\}$. Vì $B = \{TT\}$, $P(B \text{ VÀ } C) = 0$. B và C là xung khắc. (B và C không có phần tử chung nào vì bạn không thể có tất cả mặt úp và tất cả mặt ngửa cùng một lúc.)
- Gọi D = sự kiện nhận được nhiều hơn một mặt úp. $D = \{TT\}$. $P(D) = \frac{1}{4}$
- Gọi E = sự kiện nhận được mặt ngửa ở lần tung đầu tiên. (Điều này ngụ ý bạn có thể nhận được mặt ngửa hoặc mặt úp ở lần tung thứ hai.) $E = \{HT, HH\}$. $P(E) = \frac{2}{4}$

Tìm xác suất nhận được ít nhất một (một hoặc hai) mặt úp trong hai lần tung. Gọi F = sự kiện nhận được ít nhất một mặt úp trong hai lần tung. $F = \{HT, TH, TT\}$. $P(F) = \frac{3}{4}$

Bài tập vận dụng 3.6

Rút hai lá bài từ một bộ bài chuẩn 52 lá có hoàn lại. Tìm xác suất nhận được ít nhất một lá bài màu đen. Gọi E = sự kiện nhận được mặt ngửa ở lần tung đầu tiên. (Điều này ngụ ý bạn có thể nhận được mặt ngửa hoặc mặt úp ở lần tung thứ hai.) $E = \{HT, HH\}$. $P(E) = \frac{2}{4}$

Ví dụ 3.7

Vấn đề

Tung hai đồng xu công bằng. Tìm xác suất của các sự kiện.

- Gọi F = sự kiện nhận được nhiều nhất một mặt úp (không có hoặc một mặt úp).
- Gọi G = sự kiện nhận được hai mặt giống nhau.
- Gọi H = sự kiện nhận được mặt ngửa ở lần tung đầu tiên, sau đó là mặt ngửa hoặc mặt úp ở lần tung thứ hai.
- F và G có xung khắc không?
- Gọi J = sự kiện nhận được tất cả mặt úp. J và H có xung khắc không?

Lời giải

Xem không gian mẫu trong Ví dụ 3.6.

- a. Không (0) mặt úp hoặc một (1) mặt úp xuất hiện khi có các phương án HH, TH, HT.
 $P(F) = \frac{3}{4}$
- b. Hai mặt giống nhau nếu HH hoặc TT xuất hiện. $P(G) = \frac{2}{4}$
- c. Một mặt ngửa ở lần tung đầu tiên, sau đó là mặt ngửa hoặc mặt úp ở lần tung thứ hai xảy ra khi HH hoặc HT xuất hiện. $P(H) = \frac{2}{4}$
- d. F và G có chung HH nên $P(F \text{ VÀ } G)$ không bằng không (0). F và G không xung khắc.
- e. Nhận được tất cả mặt úp xảy ra khi mặt úp xuất hiện trên cả hai đồng xu (TT). Các phương án của H là HH and HT.
- J và H không có gì chung nên $P(J \text{ VÀ } H) = 0$. J và H là xung khắc.

Bài tập vận dụng 3.7

Một hộp có hai viên bi, một trắng và một đỏ. Chúng ta chọn một viên bi, đặt nó trở lại hộp, và chọn viên bi thứ hai (lấy mẫu có hoàn lại). Tìm xác suất của các sự kiện sau:

- Gọi F = sự kiện nhận được viên bi trắng hai lần.
- Gọi G = sự kiện nhận được hai viên bi có màu khác nhau.
- Gọi H = sự kiện nhận được viên trắng ở lần chọn đầu tiên.
- F và G có xung khắc không?
- G và H có xung khắc không?

Ví dụ 3.8

Tung một con xúc xắc sáu mặt cong bằng. Không gian mẫu là $\{1, 2, 3, 4, 5, 6\}$. Gọi sự kiện A = mặt là số lẻ. Khi đó $A = \{1, 3, 5\}$. Gọi sự kiện B = mặt là số chẵn. Khi đó $B = \{2, 4, 6\}$.

- Tìm phần bù của A, A'. Phần bù của A, A', là B bởi vì A và B cùng nhau tạo nên không gian mẫu. $P(A) + P(B) = P(A) + P(A') = 1$. Ngoài ra, $P(A) = \frac{3}{6}$ và $P(B) = \frac{3}{6}$.
- Gọi sự kiện C = các mặt lẻ lớn hơn 2. Khi đó $C = \{3, 5\}$. Gọi sự kiện D = tất cả các mặt chẵn nhỏ hơn 5. Khi đó $D = \{2, 4\}$. $P(C \cap D) = 0$ vì bạn không thể có một mặt lẻ và chẵn cùng một lúc. Do đó, C và D là các sự kiện xung khắc.
- Gọi sự kiện E = tất cả các mặt nhỏ hơn năm. $E = \{1, 2, 3, 4\}$.

Vấn đề

C và E có phải là các sự kiện xung khắc không? (Trả lời có hoặc không.) Tại sao?

Lời giải

Không. $C = \{3, 5\}$ và $E = \{1, 2, 3, 4\}$. $P(C \text{ AND } E) = \frac{1}{6}$. Để xung khắc, $P(C \text{ AND } E)$ phải bằng 0.

Tìm $P(C|A)$. Đây là một xác suất có điều kiện. Nhớ rằng sự kiện C là $\{3, 5\}$ và sự kiện A là $\{1, 3, 5\}$. Để tìm $P(C|A)$, tìm xác suất của C sử dụng không gian mẫu A. Bạn đã thu hẹp

không gian mẫu từ không gian mẫu ban đầu $\{1, 2, 3, 4, 5, 6\}$ xuống còn $\{1, 3, 5\}$. Vì vậy, $P(C|A) = \frac{2}{3}$.

Bài tập vận dụng 3.8

Gọi sự kiện A = học tiếng Tây Ban Nha. Gọi sự kiện B = học tiếng Đức. Khi đó A VÀ B = học tiếng Tây Ban Nha và tiếng Đức. Giả sử $P(A) = 0,4$ và $P(B) = 0,2$. $P(A$ VÀ $B) = 0,08$. Các sự kiện A và B có độc lập không? Gợi ý: Bạn phải chứng minh **MỘT** trong những điều sau:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A$ VÀ $B) = P(A)P(B)$

Ví dụ 3.9

Gọi sự kiện G = tham gia một lớp toán. Gọi sự kiện H = tham gia một lớp khoa học. Khi đó, G VÀ H = tham gia một lớp toán và một lớp khoa học. Giả sử $P(G) = 0,6$, $P(H) = 0,5$ và $P(G$ VÀ $H) = 0,3$. G và H có độc lập không?

Nếu G và H độc lập, thì bạn phải chứng minh **MỘT** trong những điều sau:

- $P(G|H) = P(G)$
- $P(H|G) = P(H)$
- $P(G$ VÀ $H) = P(G)P(H)$

Lưu ý: Lựa chọn bạn đưa ra phụ thuộc vào thông tin bạn có. Bạn có thể chọn bất kỳ phương pháp nào ở đây vì bạn có thông tin cần thiết.

Vấn đề

- Chứng minh rằng $P(G|H) = P(G)$.
- Chứng minh $P(G \cap H) = P(G)P(H)$.

Lời giải

- $P(G|H) = \frac{P(G \cap H)}{P(H)} = \frac{0,3}{0,5} = 0,6 = P(G)$
- $P(G)P(H) = (0,6)(0,5) = 0,3 = P(G \cap H)$

Vì G và H độc lập, biết rằng một người đang tham gia một lớp khoa học không làm thay đổi khả năng họ đang tham gia một lớp toán. Nếu hai sự kiện không độc lập (nghĩa là chúng phụ thuộc) thì biết rằng một người đang tham gia một lớp khoa học sẽ làm thay đổi khả năng họ đang tham gia toán. Để thực hành, hãy chứng minh rằng $P(H|G) = P(H)$ để chứng minh rằng G và H là các sự kiện độc lập.

Bài tập vận dụng 3.9

Trong một túi, có sáu viên bi đỏ và bốn viên bi xanh lá. Các viên bi đỏ được đánh dấu với các số 1, 2, 3, 4, 5 và 6. Các viên bi xanh lá được đánh dấu với các số 1, 2, 3 và 4.

- R = một viên bi đỏ
- G = một viên bi xanh
- O = một viên bi có số lẻ

Không gian mẫu là $S = \{R1, R2, R3, R4, R5, R6, G1, G2, G3, G4\}$.

S có mười phương án. $P(G \cap O)$ bằng bao nhiêu?

Ví dụ 3.10

Vấn đề

Gọi sự kiện C = tham gia một lớp tiếng Anh. Gọi sự kiện D = tham gia một lớp diễn thuyết.

Giả sử $P(C) = 0,75$, $P(D) = 0,3$, $P(C|D) = 0,75$ và $P(C \cap D) = 0,225$.

Biện minh cho câu trả lời của bạn cho các câu hỏi sau đây bằng số.

- C và D có độc lập không?
- C và D có xung khắc không?
- $P(D|C)$ bằng bao nhiêu?

Lời giải

Có, vì $P(C|D) = P(C)$.

Không, vì $P(C \cap D)$ không bằng không.

$$P(D|C) = \frac{P(C \cap D)}{P(C)} = \frac{0,225}{0,75} = 0,3$$

Bài tập vận dụng 3.10

Một sinh viên đi đến thư viện. Gọi các sự kiện B = sinh viên mượn một cuốn sách và D = sinh viên mượn một đĩa DVD. Giả sử rằng $P(B) = 0,40$, $P(D) = 0,30$ và $P(B \text{ VÀ } D) = 0,20$.

- Tìm $P(B|D)$.
- Tìm $P(D|B)$.
- B và D có độc lập không?
- B và D có xung khắc không?

Ví dụ 3.11

Trong một hộp có ba lá bài đỏ và năm lá bài xanh. Các lá bài đỏ được đánh dấu với các số 1, 2 và 3, và các lá bài xanh được đánh dấu với các số 1, 2, 3, 4 và 5. Các lá bài được xáo trộn kỹ. Bạn đưa tay vào hộp (bạn không thể nhìn vào bên trong) và rút một lá bài.

Gọi R = rút được lá bài đỏ, B = rút được lá bài xanh, E = rút được lá bài có số chẵn.

Không gian mẫu $S = R1, R2, R3, B1, B2, B3, B4, B5$. S có tám phương án.

- $P(R) = \frac{3}{8}$, $P(B) = \frac{5}{8}$, $P(R \text{ VÀ } B) = 0$. (Bạn không thể rút một lá bài vừa đỏ vừa xanh..)
- $P(E) = \frac{3}{8}$. (Có ba lá bài có số chẵn, $R2, B2$, and $B4$.)
- $P(E|B) = \frac{2}{5}$. (Có năm lá bài xanh: $B1, B2, B3, B4$, and $B5$. Trong số các lá bài xanh, có hai lá bài chẵn; $B2$ and $B4$.)
- $P(B|E) = \frac{2}{3}$. (Có ba lá bài có số chẵn: $R2, B2$, and $B4$. Trong số các lá bài có số chẵn, hai lá là xanh; $B2$ and $B4$.)
- Các sự kiện R và B là xung khắc vì $P(R \text{ VÀ } B) = 0$.

Gọi G = lá bài có số lớn hơn 3. $G = \{B4, B5\}$. $P(G) = \frac{2}{8}$. Let H = Các sự kiện R và B là xung khắc vì. $H = \{B1, B2, B3, B4\}$. $P(G|H) = \frac{1}{4}$. (Lá bài duy nhất trong H có số lớn hơn ba là $B4$.) Vì $\frac{2}{8} = \frac{1}{4}$, $P(G) = P(G|H)$, có nghĩa là G và H độc lập.

Bài tập vận dụng 3.11

Trong một sân vận động bóng rổ,

- 70% người hâm mộ đang cổ vũ cho đội nhà.
- 25% người hâm mộ đang mặc màu xanh dương.
- 20% người hâm mộ đang mặc màu xanh dương và đang cổ vũ cho đội khách.

Trong số những người hâm mộ đang cổ vũ cho đội khách, 67% đang mặc màu xanh dương.

Ví dụ 3.12

Vấn đề

Trong một lớp học đại học, 60% sinh viên là nữ. 50% tất cả sinh viên trong lớp có tóc dài. 40% sinh viên là nữ và có tóc dài. Trong số các sinh viên nữ, 75% có tóc dài. Gọi W là sự kiện một sinh viên là nữ. Gọi L là sự kiện một sinh viên có tóc dài. Một sinh viên được chọn ngẫu nhiên. Các sự kiện là nữ và có tóc dài có độc lập không?

Các xác suất sau được cho trong ví dụ này:

- $P(W) = 0,60$; $P(L) = 0,50$
- $P(W \cap L) = 0,45$

- $P(L|W) = 0,75$

Lưu ý: Lựa chọn bạn đưa ra phụ thuộc vào thông tin bạn có. Bạn có thể sử dụng điều kiện đầu tiên hoặc cuối cùng trong danh sách cho ví dụ này. Bạn chưa biết $P(W|L)$, vì vậy bạn không thể sử dụng điều kiện thứ hai.

Lời giải

Kiểm tra xem $P(W \text{ VÀ } L) = P(W)P(L)$ hay không. Chúng ta biết rằng $P(W \text{ VÀ } L) = 0,45$, nhưng $P(W)P(L) = (0,60)(0,50) = 0,30$. Các sự kiện là nữ và có tóc dài không độc lập vì $P(W \text{ VÀ } L)$ không bằng $P(W)P(L)$.

Kiểm tra xem $P(L|W)$ có bằng $P(L)$ hay không. Chúng ta biết rằng $P(L|W) = 0,75$, nhưng $P(L) = 0,50$; chúng không bằng nhau. Các sự kiện là nữ và có tóc dài không độc lập.

Các sự kiện là nữ và có tóc dài không độc lập; biết rằng một sinh viên là nữ làm thay đổi xác suất rằng một sinh viên có tóc dài.

Bài tập vận dụng 3.12

Mark cần quyết định lộ trình nào để đi làm. Các lựa chọn của anh ấy là I = đi đường cao tốc và F = đi đường số 5.

- $P(I) = 0,44$ và $P(F) = 0,56$
- $P(I \cap F) = 0$ vì Mark sẽ chỉ đi một lộ trình để đến nơi làm việc.

Xác suất của $P(I \cup F)$ là bao nhiêu?

Ví dụ 3.13

Vấn đề

- Tung một đồng xu công bằng (đồng xu có hai mặt, H và T). Các phương án là _____. Đếm các phương án. Có ____ phương án.
- Tung một con xúc xắc sáu mặt công bằng (con xúc xắc có 1, 2, 3, 4, 5 hoặc 6 chấm trên một mặt). Các phương án là _____. Đếm các phương án. Có ____ phương án.
- Nhân hai số phương án. Câu trả lời là _____.
- Nếu bạn tung một đồng xu công bằng và tiếp theo là tung một con xúc xắc sáu mặt công bằng, câu trả lời trong phần c. là số phương án (kích thước của không gian mẫu). Các phương án là gì? (Gợi ý: Hai trong số các phương án là $H1$ và $T6$.)
- Sự kiện A = mặt ngửa (H) trên đồng xu sau đó là một số chẵn (2, 4, 6) trên con xúc xắc. $A = \{ \text{_____} \}$. Tìm $P(A)$.
- Sự kiện B = mặt ngửa trên đồng xu sau đó là số ba trên con xúc xắc. $B = \{ \text{_____} \}$. Tìm $P(B)$.
- A và B có xung khắc không? (Gợi ý: $P(A \cap B)$ là bao nhiêu? Nếu $P(A \cap B) = 0$, thì A và B là xung khắc.)
- A và B có độc lập không? (Gợi ý: $P(A \cap B) = P(A)P(B)$ phải không? Nếu $P(A \cap B) = P(A)P(B)$, thì A và B độc lập. Nếu không, thì chúng phụ thuộc).

Lời giải

- a. H và T ; 2
- b. 1, 2, 3, 4, 5, 6; 6
- c. $2(6) = 12$
- d. $T1, T2, T3, T4, T5, T6, H1, H2, H3, H4, H5, H6$
- e. $A = \{H2, H4, H6\}; P(A) = \frac{3}{12}$
- f. $B = \{H3\}; P(B) = \frac{1}{12}$
- g. Có, vì $P(A \cap B) = 0$

$P(A \cap B) = 0. P(A)P(B) = \left(\frac{3}{12}\right)\left(\frac{1}{12}\right)$. $P(A \text{ VÀ } B)$ không bằng $P(A)P(B)$, vì vậy A và B phụ thuộc.

Bài tập vận dụng 3.13

Một hộp có hai viên bi, một trắng và một đỏ. Chúng ta chọn một viên bi, đặt nó trở lại hộp, và chọn viên bi thứ hai (lấy mẫu có hoàn lại). Gọi T là sự kiện nhận được viên bi trắng hai lần, F là sự kiện nhận được viên bi trắng ở lần chọn đầu tiên, S là sự kiện nhận được viên bi trắng ở lần rút thứ hai.

- Tính $P(T)$.
- Tính $P(T|F)$.
- T và F có độc lập không?
- F và S có xung khắc không?
- F và S có độc lập không?

3.3 Hai quy tắc cơ bản của xác suất

Khi tính toán xác suất, có hai quy tắc cần xem xét khi xác định liệu hai sự kiện có độc lập hay phụ thuộc và liệu chúng có xung khắc hay không.

Quy tắc nhân (multiplication rule)

Nếu A và B là hai sự kiện được định nghĩa trên một không gian mẫu, thì: $P(A \text{ VÀ } B) = P(B)P(A|B)$.

Quy tắc này cũng có thể được viết là: $P(A|B) = \frac{P(A \cap B)}{P(B)}$

(Xác suất của A nếu có B bằng xác suất của A và B chia cho xác suất của B .)

Nếu A và B độc lập, thì $P(A|B) = P(A)$. Khi đó $P(A \cap B) = P(A|B)P(B)$ trở thành $P(A \cap B) = P(A)P(B)$.

Quy tắc cộng (addition rule)

Nếu A và B được định nghĩa trên một không gian mẫu, thì: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Nếu A và B xung khắc, thì $P(A \text{ VÀ } B) = 0$. Khi đó $P(A \text{ HOẶC } B) = P(A) + P(B) - P(A \cap B)$ trở thành $P(A \cup B) = P(A) + P(B)$.

Ví dụ 3.14

Klaus đang cố gắng chọn nơi để đi nghỉ. Hai lựa chọn của anh ấy là: A = đi New Zealand và B = đi Alaska

Klaus chỉ có thể chi trả cho một kỳ nghỉ. Xác suất rằng anh ấy chọn A là $P(A) = 0,6$ và xác suất rằng anh ấy chọn B là $P(B) = 0,35$.

$P(A \cap B) = 0$ vì Klaus chỉ có thể chi trả để đi một kỳ nghỉ

Do đó, xác suất rằng anh ấy chọn New Zealand hoặc Alaska là $P(A \cup B) = P(A) + P(B) = 0,6 + 0,35 = 0,95$. Lưu ý rằng xác suất anh ấy không chọn đi bất kỳ đâu nghỉ phải là 0,05.

Bài tập vận dụng 3.14

Anna phải mua một chiếc xe mới. Cô ấy có hai lựa chọn, xe A và xe B. Anna chỉ có thể mua một chiếc xe. Xác suất Anna sẽ mua xe A là $P(A) = 0,25$, và xác suất Anna sẽ mua xe B là $P(B) = 0,65$. Tìm:

- $P(A \text{ VÀ } B)$
- $P(A \text{ HOẶC } B)$

Ví dụ 3.15

Carlos là cầu thủ bóng đá. Anh ấy ghi bàn 65% thời gian anh ấy sút. Carlos cố gắng ghi hai bàn liên tiếp trong trận đấu tiếp theo. A = sự kiện Carlos thành công trong lần thử đầu tiên của anh ấy. $P(A) = 0,65$. B = sự kiện Carlos thành công trong lần thử thứ hai của anh ấy. $P(B) = 0,65$. Carlos có xu hướng sút theo chuỗi. Xác suất anh ấy ghi bàn thứ hai **VỚI ĐIỀU KIỆN** anh ấy đã ghi bàn đầu tiên là 0,90.

Vấn đề

- Xác suất anh ấy ghi cả hai bàn là bao nhiêu?
- Xác suất Carlos ghi bàn đầu tiên hoặc bàn thứ hai là bao nhiêu?
- A và B có độc lập không?
- A và B có xung khắc không?

Lời giải

a. Bài toán yêu cầu bạn tìm $P(A \text{ VÀ } B) = P(B \text{ VÀ } A)$. Vì $P(B|A) = 0,90$: $P(B \text{ VÀ } A) = P(B|A)P(A) = (0,90)(0,65) = 0,585$

Carlos ghi bàn thứ nhất và thứ hai với xác suất 0,585.

b. Bài toán yêu cầu bạn tìm $P(A \text{ HOẶC } B)$.

$$P(A \text{ HOẶC } B) = P(A) + P(B) - P(A \text{ VÀ } B) = 0,65 + 0,65 - 0,585 = 0,715$$

Carlos ghi bàn thứ nhất hoặc bàn thứ hai với xác suất 0,715.

c. Không, chúng không độc lập, vì $P(B \text{ VÀ } A) = 0,585$.

$$P(B)P(A) = (0,65)(0,65) = 0,423$$

$$0,423 \neq 0,585 = P(B \text{ VÀ } A)$$

Vì vậy, $P(B \text{ VÀ } A)$ **không** bằng $P(B)P(A)$.

d. Không, chúng không xung khắc vì $P(A \text{ và } B) = 0,585$.

Để xung khắc, $P(A \text{ VÀ } B)$ phải bằng không.

Bài tập vận dụng 3.15

Helen chơi bóng rổ. Khi ném phạt, cô ấy thực hiện cú ném thành công 75% thời gian. Helen bây giờ phải thực hiện hai cú ném phạt. C = sự kiện Helen thực hiện thành công cú ném đầu tiên. $P(C) = 0,75$. D = sự kiện Helen thực hiện thành công cú ném thứ hai. $P(D) = 0,75$. Xác suất Helen thực hiện thành công cú ném phạt thứ hai với điều kiện cô ấy đã thực hiện thành công cú ném đầu tiên là 0,85. Xác suất Helen thực hiện thành công cả hai cú ném phạt là bao nhiêu?

Ví dụ 3.16

Một đội bơi cộng đồng có **150** thành viên. 75 thành viên là những người bơi nâng cao. **Bốn mươi bảy** thành viên là những người bơi trung cấp. Số còn lại là những người bơi sơ cấp. **Bốn mươi** người bơi nâng cao tập luyện bốn lần một tuần. **Ba mươi** người bơi trung cấp tập luyện bốn lần một tuần. **Mười** người bơi sơ cấp tập luyện bốn lần một tuần. Giả sử một thành viên của đội bơi được chọn ngẫu nhiên.

Bài toán

- Xác suất thành viên đó là người bơi sơ cấp là bao nhiêu?
- Xác suất thành viên đó tập luyện bốn lần một tuần là bao nhiêu?
- Xác suất thành viên đó là người bơi nâng cao và tập luyện bốn lần một tuần là bao nhiêu?
- Xác suất một thành viên là người bơi nâng cao và người bơi trung cấp là bao nhiêu? Việc là người bơi nâng cao và người bơi trung cấp có xung khắc không? Tại sao?
- Việc là người bơi sơ cấp và tập luyện bốn lần một tuần có phải là các sự kiện độc lập không? Tại sao?

Lời giải

a. $\frac{28}{150}$

b. $\frac{80}{150}$

c. $\frac{40}{150}$

d. $P(\text{nâng cao VÀ trung cấp}) = 0$, vì vậy đây là các sự kiện xung khắc. Một người bơi không thể vừa là người bơi nâng cao vừa là người bơi trung cấp cùng một lúc.

e. Không, đây không phải là các sự kiện độc lập.

$$P(\text{sơ cấp VÀ tập luyện bốn lần mỗi tuần}) = 0,0667$$

$$P(\text{sơ cấp})P(\text{tập luyện bốn lần mỗi tuần}) = 0,0996$$

$$0,0667 \neq 0,0996$$

Bài tập vận dụng 3.16

Một trường học có 200 học sinh năm cuối, trong đó 140 người sẽ học đại học vào năm tới. 40 người sẽ đi làm trực tiếp. Số còn lại đang có 5 nghỉ ngơi. Năm mươi học sinh năm cuối đi học đại học chơi thể thao. Ba mươi học sinh năm cuối đi làm trực tiếp chơi thể thao. Năm học sinh năm cuối có năm nghỉ ngơi chơi thể thao. Xác suất một học sinh năm cuối đang có năm nghỉ ngơi là bao nhiêu?

Ví dụ 3.17

Felicity học tại Modesto JC ở Modesto, CA. Xác suất Felicity đăng ký một lớp toán là 0,2 và xác suất cô ấy đăng ký một lớp diễn thuyết là 0,65. Xác suất cô ấy đăng ký một lớp toán VỚI ĐIỀU KIỆN cô ấy đăng ký lớp diễn thuyết là 0,25.

Gọi: M = lớp toán, S = lớp diễn thuyết, $M|S$ = toán với điều kiện diễn thuyết

Vấn đề

a. Xác suất Felicity đăng ký toán và diễn thuyết là bao nhiêu? Tìm $P(M \text{ VÀ } S) = P(M|S)P(S)$.

b. Xác suất Felicity đăng ký lớp toán hoặc lớp diễn thuyết là bao nhiêu? Tìm $P(M \text{ HOẶC } S) = P(M) + P(S) - P(M \text{ VÀ } S)$.

c. M và S có độc lập không? $P(M|S) = P(M)$ phải không?

d. M và S có xung khắc không? $P(M \text{ VÀ } S) = 0$ phải không?

Lời giải

a. 0,1625

b. 0,6875

c. Không

d. Không

Bài tập vận dụng 3.17

Một sinh viên đi đến thư viện. Gọi các sự kiện B = sinh viên mượn một cuốn sách và D = sinh viên mượn một đĩa DVD. Giả sử rằng $P(B) = 0,40$, $P(D) = 0,30$ và $P(D|B) = 0,5$.

- Tìm $P(B \text{ VÀ } D)$.
- Tìm $P(B \text{ HOẶC } D)$.

Ví dụ 3.18

Các nghiên cứu cho thấy khoảng 1 phụ nữ trong 7 (khoảng 14,3%) sống đến 90 tuổi sẽ phát triển ung thư vú. Giả sử rằng trong số những phụ nữ phát triển ung thư vú, một xét nghiệm âm tính 2% thời gian. Ngoài ra, giả sử rằng trong dân số phụ nữ nói chung, xét nghiệm ung thư vú âm tính khoảng 85% thời gian. Gọi B = phụ nữ phát triển ung thư vú và gọi N = xét nghiệm âm tính. Giả sử một phụ nữ được chọn ngẫu nhiên.

Vấn đề

- Xác suất người phụ nữ đó phát triển ung thư vú là bao nhiêu? Xác suất người phụ nữ đó xét nghiệm âm tính là bao nhiêu?
- Với điều kiện người phụ nữ đó bị ung thư vú, xác suất cô ấy xét nghiệm âm tính là bao nhiêu?
- Xác suất người phụ nữ đó bị ung thư vú VÀ xét nghiệm âm tính là bao nhiêu?
- Xác suất người phụ nữ đó bị ung thư vú hoặc xét nghiệm âm tính là bao nhiêu?
- Việc bị ung thư vú và xét nghiệm âm tính có phải là các sự kiện độc lập không?
- Việc bị ung thư vú và xét nghiệm âm tính có xung khắc không?

Lời giải

- $P(B) = 0,143$; $P(N) = 0,85$
- $P(N|B) = 0,02$
- $P(B \text{ VÀ } N) = P(B)P(N|B) = (0,143)(0,02) = 0,0029$
- $P(B \text{ HOẶC } N) = P(B) + P(N) - P(B \text{ VÀ } N) = 0,143 + 0,85 - 0,0029 = 0,9901$
- Không. $P(N) = 0,85$; $P(N|B) = 0,02$. Vì vậy, $P(N|B)$ không bằng $P(N)$.
- Không. $P(B \text{ VÀ } N) = 0,0029$. Để B và N xung khắc, $P(B \text{ VÀ } N)$ phải bằng không.

Bài tập vận dụng 3.18

Một trường học có 200 học sinh năm cuối, trong đó 140 người sẽ học đại học vào năm tới. 40 người sẽ đi làm trực tiếp. Số còn lại đang có năm nghỉ ngơi. Năm mươi học sinh năm cuối đi học đại học chơi thể thao. Ba mươi học sinh năm cuối đi làm trực tiếp chơi thể thao. Năm học sinh năm cuối có năm nghỉ ngơi chơi thể thao. Xác suất một học sinh năm cuối đi học đại học và chơi thể thao là bao nhiêu?

Ví dụ 3.19*Vấn đề*

Tham khảo thông tin trong Ví dụ 3.18. P = xét nghiệm dương tính.

- Với điều kiện một phụ nữ phát triển ung thư vú, xác suất cô ấy xét nghiệm dương tính là bao nhiêu. Tìm $P(P|B) = 1 - P(N|B)$.
- Xác suất một phụ nữ phát triển ung thư vú và xét nghiệm dương tính là bao nhiêu. Tìm $P(B \text{ VÀ } P) = P(P|B)P(B)$.
- Xác suất một phụ nữ không phát triển ung thư vú là bao nhiêu. Tìm $P(B') = 1 - P(B)$.
- Xác suất một phụ nữ xét nghiệm dương tính đối với ung thư vú là bao nhiêu. Tìm $P(P) = 1 - P(N)$.

Lời giải

- 0,98
- 0,1401
- 0,857
- 0,15

Bài tập vận dụng 3.19

Một sinh viên đi đến thư viện. Gọi các sự kiện B = sinh viên mượn một cuốn sách và D = sinh viên mượn một đĩa DVD. Giả sử rằng $P(B) = 0,40$, $P(D) = 0,30$ và $P(D|B) = 0,5$.

- Tìm $P(B')$.
- Tìm $P(D \text{ VÀ } B)$.
- Tìm $P(B|D)$.
- Tìm $P(D \text{ VÀ } B')$.
- Tìm $P(D|B')$.

3.4 Bảng tiếp liên

Bảng tiếp liên⁹ (contingency table) cung cấp một cách thể hiện dữ liệu giúp tính toán xác suất. Bảng này giúp xác định các xác suất có điều kiện khá dễ dàng. Bảng hiển thị các giá trị mẫu trong mối quan hệ với hai biến khác nhau có thể phụ thuộc lẫn nhau. Phần sau, chúng ta sẽ sử dụng bảng tiếp liên một lần nữa, nhưng theo cách khác.

⁹ Còn gọi là “bảng tương quan”

Ví dụ 3.20*Bài toán*

Giả sử ta có một nghiên cứu về vi phạm tốc độ và người lái xe sử dụng điện thoại di động với các dữ liệu hư cấu như sau:

Bảng 3.3.

	Vi phạm tốc độ trong năm qua	Không vi phạm tốc độ trong năm qua	Tổng
Sử dụng điện thoại khi lái xe	25	280	305
Không sử dụng điện thoại khi lái xe	45	405	450
Tổng	70	685	755

Tổng số người trong mẫu là 755. Tổng các hàng là 305 và 450. Tổng các cột là 70 và 685. Lưu ý rằng $305 + 450 = 755$ và $70 + 685 = 755$.

Tính các xác suất sau sử dụng bảng.

- Tìm $P(\text{Người lái xe là người sử dụng điện thoại di động})$.
- Tìm $P(\text{người lái xe không có vi phạm trong năm ngoái})$.
- Tìm $P(\text{Người lái xe không có vi phạm trong năm ngoái VÀ là người sử dụng điện thoại di động})$.
- Tìm $P(\text{Người lái xe là người sử dụng điện thoại di động HOẶC người lái xe không có vi phạm trong năm ngoái})$.
- Tìm $P(\text{Người lái xe là người sử dụng điện thoại di động VỚI ĐIỀU KIỆN người lái xe có vi phạm trong năm ngoái})$.
- Tìm $P(\text{Người lái xe không có vi phạm năm ngoái VỚI ĐIỀU KIỆN người lái xe không phải là người sử dụng điện thoại di động})$.

Lời giải

$$\text{a. } \frac{\text{số người sử dụng điện thoại di động}}{\text{tổng số người trong nghiên cứu}} = \frac{305}{755}$$

$$\text{b. } \frac{\text{tổng số người không có vi phạm}}{\text{tổng số người trong nghiên cứu}} = \frac{685}{755}$$

$$\text{c. } \frac{280}{755}$$

$$\text{d. } \left(\frac{305}{755} + \frac{685}{755} \right) - \frac{280}{755} = \frac{710}{755}$$

$$\text{e. } \frac{25}{70} \text{ (Không gian mẫu được thu hẹp thành số người lái xe có vi phạm.)}$$

$$\text{f. } \frac{405}{450} \text{ (Không gian mẫu được thu hẹp thành số người lái xe không phải là người sử dụng điện thoại di động.)}$$

Bài tập vận dụng 3.20

Bảng 3.4 cho thấy số lượng vận động viên thực hiện làm nóng trước khi tập thể dục và số lượng bị chấn thương trong năm qua.

Bảng 3.4

	Chấn thương trong năm ngoái	Không chấn thương trong năm ngoái	Tổng cộng
Làm nóng	55	295	350
Không làm nóng	231	219	450
Tổng cộng	286	514	800

- a. $P(\text{vận động viên làm nóng trước khi tập thể dục})$ là bao nhiêu?
- b. $P(\text{vận động viên làm nóng trước khi tập thể dục} | \text{không chấn thương trong năm ngoái})$ là bao nhiêu?

Ví dụ 3.21

Bảng 3.5 cho thấy một mẫu ngẫu nhiên của 100 người đi bộ đường dài và các khu vực đi bộ họ ưa thích.

Bảng 3.5 Sở thích khu vực đi bộ đường dài

Giới tính	Bờ biển	Gần hồ và suối	Trên đỉnh núi	Tổng cộng
Nữ	18	16		45
Nam			14	55
Tổng cộng		41		

Vấn đề

- a. Hoàn chỉnh các chỗ còn trống trong bảng
- b. Các sự kiện "là phụ nữ" và "thích bờ biển" có phải là các sự kiện độc lập không?

Gọi F = là phụ nữ và gọi C = thích bờ biển.

Tìm $P(F \text{ VÀ } C)$.

Tìm $P(F)P(C)$

Hai số này có bằng nhau không? Nếu chúng bằng nhau, thì F và C độc lập. Nếu chúng không bằng nhau, thì F và C không độc lập.

- c. Tìm xác suất một người là nam nếu người đó thích đi bộ đường dài gần hồ và suối. Gọi M = là nam, và gọi L = thích đi bộ đường dài gần hồ và suối.

- Từ nào cho bạn biết đây là một điều kiện?

- Điền vào chỗ trống và tính xác suất: $P(_\mid_) = _\$.
- Không gian mẫu cho câu hỏi này có phải là tất cả 100 người đi bộ đường dài không? Nếu không, nó là gì?

d. Tìm xác suất một người là phụ nữ, hoặc là người thích đi bộ đường dài trên đỉnh núi. Gọi F = là phụ nữ, và gọi P = thích đỉnh núi.

- Tìm $P(F)$.
- Tìm $P(P)$.
- Tìm $P(F \text{ VÀ } P)$.
- Tìm $P(F \text{ HOẶC } P)$.

Lời giải

a.

Bảng 3.6 Sở thích khu vực đi bộ đường dài

Giới tính	Bờ biển	Gần hồ và suối	Trên đỉnh núi	Tổng cộng
Nữ	18	16	11	45
Nam	16	25	14	55
Tổng cộng	34	41	25	

b.

- $P(F \text{ VÀ } C) = \frac{18}{100} = 0.18$
- $P(F)P(C) = \left(\frac{45}{100}\right)\left(\frac{34}{100}\right) = (0,45)(0,34) = 0,153$
- $P(F \text{ VÀ } C) \neq P(F)P(C)$, vì vậy các sự kiện F và C không độc lập.

c.

- Từ “nếu” cho bạn biết đây là một điều kiện.
- $P(M|L) = \frac{15}{41}$
- Không, không gian mẫu cho bài toán này là 41 người đi bộ đường dài, thích hồ và suối.

d.

- $P(F) = \frac{45}{100}$
- $P(P) = \frac{25}{100}$
- $P(F \text{ VÀ } P) = \frac{11}{100}$
- $P(F \text{ HOẶC } P) = \frac{45}{100} + \frac{25}{100} - \frac{11}{100} = \frac{59}{100}$

Bài tập vận dụng 3.21

Bảng 3.7 cung cấp một mẫu ngẫu nhiên của 200 người đi xe đạp và các tuyến đường họ ưa thích. Gọi O = người lớn tuổi hơn và H = đường đồi.

Bảng 3.7

Nhóm tuổi	Đường hồ	Đường đồi	Đường rừng	Tổng cộng
Người trẻ tuổi	45	38	27	110
Người lớn tuổi	26	52	12	90
Tổng cộng	71	90	39	200

- Trong nhóm người lớn tuổi, xác suất người đi xe đạp thích đường đồi là bao nhiêu?
- Các sự kiện "người lớn tuổi" và "thích đường đồi" có phải là các sự kiện độc lập không?

Ví dụ 3.22

Chuột Muddy sống trong một lồng có ba cửa. Nếu Muddy đi ra cửa thứ nhất, xác suất chú chuột bị mèo Alissa bắt là $\frac{1}{5}$ và xác suất không bị bắt là $\frac{4}{5}$. Nếu đi ra cửa thứ hai, xác suất bị Alissa bắt là $\frac{1}{4}$ và xác suất không bị bắt là $\frac{3}{4}$. Xác suất Alissa bắt được Muddy khi ra khỏi cửa thứ ba là $\frac{1}{2}$ và xác suất không bắt được Muddy là $\frac{1}{2}$. Muddy có khả năng chọn bất kỳ cửa nào trong ba cửa là như nhau, nên xác suất chọn mỗi cửa là $\frac{1}{3}$.

Bảng 3.8 Lựa chọn cửa

Bị bắt hoặc không	Cửa 1	Cửa 2	Cửa 3	Tổng cộng
Bị bắt	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	
Không bị bắt	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	
Tổng cộng				1

Ô bảng đầu tiên là $\frac{1}{15} = \left(\frac{1}{5}\right)\left(\frac{1}{3}\right)$ là $P(\text{Cửa Một VÀ Bị bắt})$

Ô có trị $\frac{4}{15} = \left(\frac{4}{5}\right)\left(\frac{1}{3}\right)$ là $P(\text{Cửa Một VÀ Không bị bắt})$

Kiểm tra các ô còn lại.

Vấn đề

- Hoàn thành bảng tiếp liên. Tính các ô cho tổng cộng. Xác nhận rằng ô ở góc dưới bên phải là 1.
- Xác suất Alissa không bắt được Muddy là bao nhiêu?
- Xác suất Muddy chọn Cửa Một HOẶC Cửa Hai nếu Muddy bị Alissa bắt là bao nhiêu?

Lời giải

a.

Bảng 3.9 Lựa chọn cửa

Bị bắt hoặc không	Cửa 1	Cửa 2	Cửa 3	Tổng cộng
Bị bắt	$\frac{1}{15}$	$\frac{1}{12}$	$\frac{1}{6}$	$\frac{19}{60}$
Không bị bắt	$\frac{4}{15}$	$\frac{3}{12}$	$\frac{1}{6}$	$\frac{41}{60}$
Tổng cộng	$\frac{5}{15}$	$\frac{4}{12}$	$\frac{2}{6}$	1

b. $\frac{41}{60}$

c. $\frac{9}{19}$

Bài tập vận dụng 3.22

Anna phải mua một chiếc xe mới. Cô ấy có hai lựa chọn, xe A và xe B. Anna chỉ có thể mua một chiếc xe. Xác suất Anna sẽ mua xe A là $P(A) = 0,25$, và xác suất Anna sẽ mua xe B là $P(B) = 0,65$. Tìm:

- $P(A \text{ VÀ } B)$
- $P(A \text{ HOẶC } B)$

Ví dụ 3.23

Bảng 3.10 cung cấp số liệu thống kê số lượng tội phạm trên 100.000 dân ở Hoa Kỳ theo năm.

Bảng 3.10 Tỷ lệ chỉ số tội phạm Hoa Kỳ trên 100.000 dân

Năm	Cướp trộm	Phá hoại	Trộm	Phương tiện	Tổng cộng
1	145,7	732,1	29,7	314,7	
2	133,1	717,7	29,1	259,2	
3	119,3	701	27,7	239,1	
4	113,7	702,2	26,8	229,6	
Tổng cộng					

Vấn đề

TỔNG mỗi cột và mỗi hàng. Tổng dữ liệu = 4.520,7

- Tìm $P(\text{Năm 2 VÀ Cướp})$.
- Tìm $P(\text{Năm 3 VÀ Trộm đột nhập})$.
- Tìm $P(\text{Năm 3 HOẶC Trộm đột nhập})$.
- Tìm $P(\text{Năm 4} \mid \text{Phá hoại})$.
- Tìm $P(\text{Trộm phương tiện} \mid \text{Năm 1})$.

Lời giải

a. 0,0294, b. 0,1551, c. 0,7165, d. 0,2365, e. 0,2575

Bài tập vận dụng 3.23

Bảng 3.11 cung cấp sự liên quan giữa cân nặng và chiều cao của một nhóm tham gia vào một nghiên cứu quan sát.

Bảng 3.11

Cân nặng/Chiều cao	Cao	Trung bình	Thấp	Tổng
Thừa cân	18	28	14	
Bình thường	20	51	28	
Thiếu cân	12	25	9	
Tổng cộng				

- Tìm tổng cho mỗi hàng và cột
- Tìm xác suất một cá nhân được chọn ngẫu nhiên từ nhóm này là Cao.
- Tìm xác suất một cá nhân được chọn ngẫu nhiên từ nhóm này là Thừa cân và Cao.
- Tìm xác suất một cá nhân được chọn ngẫu nhiên từ nhóm này là Cao với điều kiện cá nhân đó là Thừa cân.
- Tìm xác suất một cá nhân được chọn ngẫu nhiên từ nhóm này là Thừa cân với điều kiện cá nhân đó là Cao.
- Tìm xác suất một cá nhân được chọn ngẫu nhiên từ nhóm này là Cao và Thiếu cân.
- Các sự kiện Thừa cân và Cao có độc lập không?

Sơ đồ cây

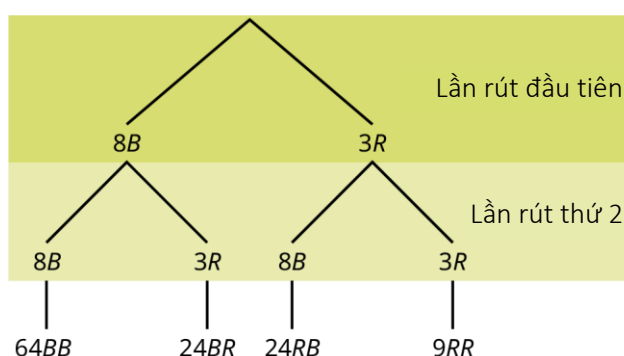
Đôi khi, khi các bài toán xác suất phức tạp, việc vẽ biểu đồ tình huống có thể hữu ích. Sơ đồ cây là công cụ có thể được sử dụng để hình dung và giải quyết các xác suất có điều kiện.

Sơ đồ cây

Sơ đồ cây (tree diagram) là một loại biểu đồ đặc biệt được sử dụng để xác định các kết quả của một thực nghiệm. Nó bao gồm các "nhánh" được ghi nhãn bằng tần suất hoặc xác suất. Sơ đồ cây có thể giúp một số bài toán xác suất dễ hình dung và giải quyết hơn. Ví dụ sau minh họa cách sử dụng sơ đồ cây.

Ví dụ 3.24

Trong một bình, có 11 quả bóng. Ba quả bóng màu đỏ (R) và tám quả bóng màu xanh (B). Rút hai quả bóng, mỗi lần một quả, **có hoàn lại**. "Có hoàn lại" có nghĩa là bạn đặt quả bóng đầu tiên trở lại bình trước khi bạn chọn quả bóng thứ hai. Dưới đây là sơ đồ cây sử dụng tần suất thể hiện tất cả các phương án có thể có.



Hình 3.2 Tổng = $64 + 24 + 24 + 9 = 121$

Tập hợp nhánh đầu tiên biểu diễn lần rút đầu tiên. Tập hợp nhánh thứ hai biểu diễn lần rút thứ hai. Mỗi phương án đều riêng biệt. Thực tế, chúng ta có thể liệt kê mỗi quả bóng đỏ là $R1, R2$ và $R3$ và mỗi quả bóng xanh là $B1, B2, B3, B4, B5, B6, B7$ và $B8$. Sau đó, 9 phương án RR có thể được viết là:

$R1R1; R1R2; R1R3; R2R1; R2R2; R2R3; R3R1; R3R2; R3R3$

Các phương án khác cũng tương tự.

Có tổng cộng 11 quả bóng trong bình. Rút hai quả bóng, mỗi lần một quả, có hoàn lại. Có $11(11) = 121$ phương án, kích thước của không gian mẫu.

Vấn đề

- Liệt kê 24 phương án BR : $B1R1, B1R2, B1R3, \dots$
- Sử dụng sơ đồ cây, tính $P(RR)$.
- Sử dụng sơ đồ cây, tính $P(RB \text{ HOẶC } BR)$.
- Sử dụng sơ đồ cây, tính $P(R \text{ ở lần rút thứ 1 VÀ } B \text{ ở lần rút thứ 2})$.
- Sử dụng sơ đồ cây, tính $P(R \text{ ở lần rút thứ 2 VỚI ĐIỀU KIỆN } B \text{ ở lần rút thứ 1})$.
- Sử dụng sơ đồ cây, tính $P(BB)$.
- Sử dụng sơ đồ cây, tính $P(B \text{ ở lần rút thứ 2 với điều kiện } R \text{ ở lần rút thứ nhất})$.

Lời giải

a. $B1R1; B1R2; B1R3; B2R1; B2R2; B2R3; B3R1; B3R2; B3R3; B4R1; B4R2; B4R3; B5R1; B5R2; B5R3; B6R1; B6R2; B6R3; B7R1; B7R2; B7R3; B8R1; B8R2; B8R3$

b. $P(RR) = \left(\frac{3}{11}\right)\left(\frac{3}{11}\right) = \frac{9}{121}$

c. $P(RB \text{ HOẶC } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{11}\right) = \frac{48}{121}$

d. $P(R \text{ ở lần rút thứ 1 AND } B \text{ ở lần rút thứ 2}) = P(RB) = \left(\frac{3}{11}\right)\left(\frac{8}{11}\right) = \frac{24}{121}$

e. $P(R \text{ ở lần rút thứ 2 VỚI ĐIỀU KIỆN } B \text{ ở lần rút thứ 1}) = P(R \text{ ở lần rút thứ 2} | B \text{ ở lần rút thứ 1}) = \frac{24}{88} = \frac{3}{11}$

Bài toán này là bài toán có điều kiện. Không gian mẫu đã được thu hẹp thành những phương án đã có màu xanh ở lần rút đầu tiên. Có $24 + 64 = 88$ phương án có thể có (24 BR và 64 BB). Hai mươi bốn trong số 88 phương án có thể có là BR . $\frac{24}{88} = \frac{3}{11}$.

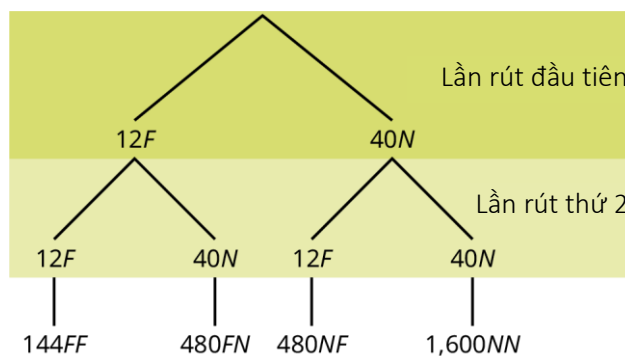
f. $P(BB) = \frac{64}{121}$

g. $P(B \text{ ở lần rút thứ 2} | R \text{ ở lần rút thứ 1}) = \frac{8}{11}$

Có $9 + 24$ phương án có R ở lần rút đầu tiên (9 RR và 24 RB). Không gian mẫu khi đó là $9 + 24 = 33$. 24 trong số 33 phương án có B ở lần rút thứ hai. Xác suất khi đó là $\frac{24}{33}$

Bài tập vận dụng 3.24

Trong một bộ bài tây, có 52 lá bài. 12 lá bài là bài hình (sự kiện F) và 40 lá bài không phải bài hình (sự kiện N). Rút 2 lá bài, mỗi lần 1 lá, có hoàn lại. Tất cả các phương án có thể có được thể hiện trong sơ đồ cây dạng tần suất. Sử dụng sơ đồ cây, tính $P(FF)$.

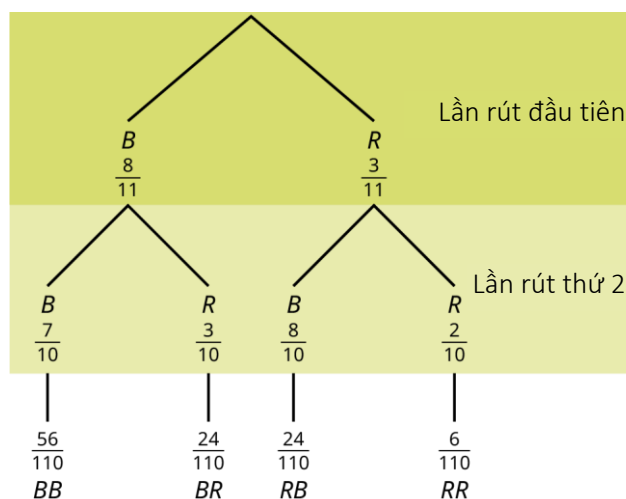


Hình 3.3

Ví dụ 3.25

Một bình có ba viên bi đỏ và tám viên bi xanh. Rút hai viên bi, mỗi lần một viên, lần này không hoàn lại, từ bình. "**Không hoàn lại**" có nghĩa là bạn không đặt quả bóng đầu tiên trở lại trước khi bạn chọn viên bi thứ hai. Sau đây là sơ đồ cây cho tình huống

này. Các nhánh được ghi nhãn bằng xác suất thay vì tần suất. Các số ở cuối các nhánh được tính bằng cách nhân các số trên hai nhánh tương ứng, ví dụ, $\left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$.



Hình 3.4 Tổng = $\frac{56+24+24+6}{110} = \frac{110}{110} = 1$

Tính các xác suất sau sử dụng sơ đồ cây.

Vấn đề

a. $P(RR) = \underline{\hspace{2cm}}$

b. Điền vào chỗ trống:

$$P(RB \text{ HOẶC } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + (\underline{\hspace{1cm}})(\underline{\hspace{1cm}}) = \frac{48}{110}$$

c. $P(R \text{ ở lần 2} | B \text{ ở lần 1}) =$

d. Điền vào chỗ trống.

$$P(R \text{ ở lần rút 1 VÀ } B \text{ ở lần rút 2}) = P(RB) = (\underline{\hspace{1cm}})(\underline{\hspace{1cm}}) = \frac{24}{110}$$

e. Tìm $P(BB)$.

f. Tìm $P(B \text{ ở lần rút 2} | R \text{ ở lần rút 1})$.

Lời giải

a. $P(RR) = \left(\frac{3}{11}\right)\left(\frac{2}{10}\right) = \frac{6}{110}$

b. $P(RB \text{ HOẶC } BR) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) + \left(\frac{8}{11}\right)\left(\frac{3}{10}\right) = \frac{48}{110}$

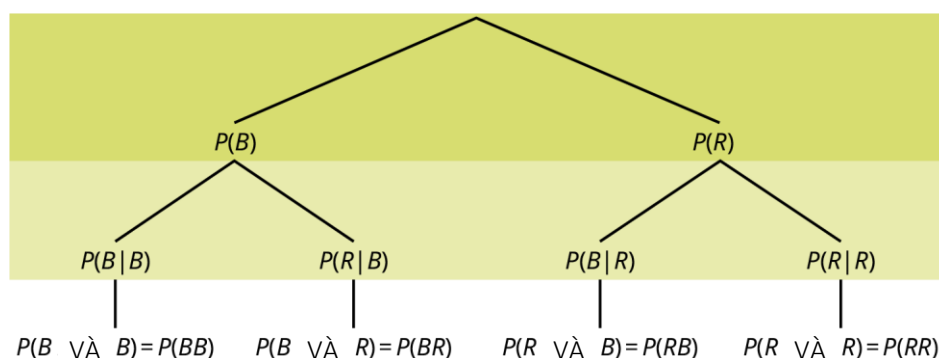
c. $P(R \text{ ở lần 2} | B \text{ ở lần 1}) = \frac{3}{10}$

d. $P(R \text{ ở lần rút 1 VÀ } B \text{ ở lần rút 2}) = P(RB) = \left(\frac{3}{11}\right)\left(\frac{8}{10}\right) = \frac{24}{110}$

e. $P(BB) = \left(\frac{8}{11}\right)\left(\frac{7}{10}\right)$

f. Sử dụng sơ đồ cây, $P(B \text{ ở lần rút 2} | R \text{ ở lần rút 1}) = P(R | B) = \frac{8}{10}$.

Nếu chúng ta sử dụng xác suất, chúng ta có thể ghi nhãn cây theo cách chung như sau.



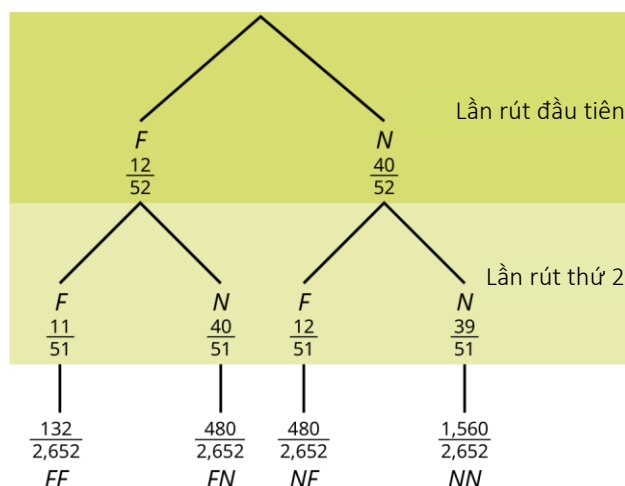
- $P(R|R)$ ở đây có nghĩa là $P(R$ ở lần rút 2 $| R$ ở lần rút 1)
- $P(B|R)$ ở đây có nghĩa là $P(B$ ở lần rút 2 $| R$ ở lần rút 1)
- $P(R|B)$ ở đây có nghĩa là $P(R$ ở lần rút 2 $| B$ ở lần rút 1)
- $P(B|B)$ ở đây có nghĩa là $P(B$ ở lần rút 2 $| B$ ở lần rút 1)

LƯU Ý

Nếu bạn rút được màu đỏ ở lần rút đầu tiên từ ba khả năng màu đỏ, còn lại hai viên bi đỏ để rút ở lần rút thứ hai. Bạn không đặt lại hoặc thay thế viên bi đầu tiên sau khi đã rút nó. Bạn rút không hoàn lại, để ở lần rút thứ hai còn lại mười viên bi trong bình.

Bài tập vận dụng 3.25

Trong một bộ bài tây, có 52 lá bài. Mười hai lá bài là bài hình (F) và 40 lá bài không phải bài hình (N). Rút hai lá bài, mỗi lần một lá, không hoàn lại. Sơ đồ cây được ghi nhãn với tất cả các xác suất có thể có.

**Hình 3.5**

- Tìm $P(FN \text{ HOẶC } NF)$.
- Tìm $P(N|F)$.

c. Tìm P (nhiều nhất một bài mặt).

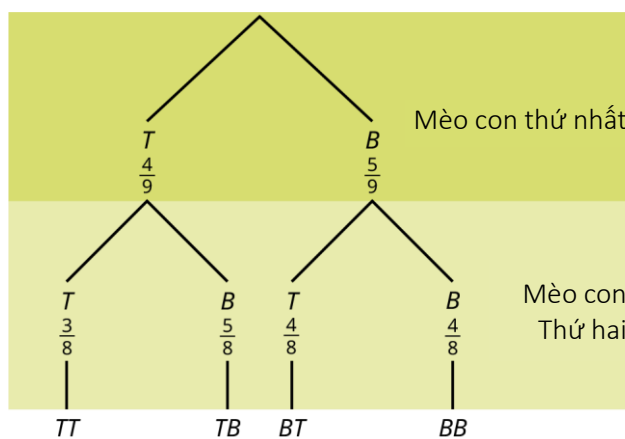
Gợi ý: "Nhiều nhất một bài mặt" có nghĩa là không có hoặc một bài mặt.

d. Tìm P (ít nhất một bài mặt).

Gợi ý: "Ít nhất một bài mặt" có nghĩa là một hoặc hai bài mặt.

Bài tập vận dụng 3.26

Một ổ mèo con có sẵn để nhận nuôi tại Hiệp hội Nhân đạo có bốn mèo con vằn và năm mèo con đen. Một gia đình đến và ngẫu nhiên chọn hai mèo con (không hoàn lại) để nhận nuôi.



Vấn đề

a. Xác suất cả hai mèo con đều là mèo vằn là bao nhiêu?

a. $\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{4}{9}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{3}{8}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$

b. Xác suất chọn được một mèo con của mỗi màu là bao nhiêu?

a. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right)$ b. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right)$ c. $\left(\frac{4}{9}\right)\left(\frac{5}{9}\right) + \left(\frac{5}{9}\right)\left(\frac{4}{9}\right)$ d. $\left(\frac{4}{9}\right)\left(\frac{5}{8}\right) + \left(\frac{5}{9}\right)\left(\frac{4}{8}\right)$

c. Xác suất mèo vằn được chọn là mèo con thứ hai khi một mèo đen được chọn là mèo con đầu tiên là bao nhiêu?

d. Xác suất chọn hai mèo con cùng màu là bao nhiêu?

Lời giải

a. c

b. d

c. $\frac{4}{8}$

d. $\frac{32}{72}$

Bài tập vận dụng 3.26

Giả sử có bốn quả bóng đỏ và ba quả bóng vàng trong một hộp. Hai quả bóng được rút ra từ hộp không hoàn lại. Xác suất chọn được 2 quả bóng khác màu là bao nhiêu?

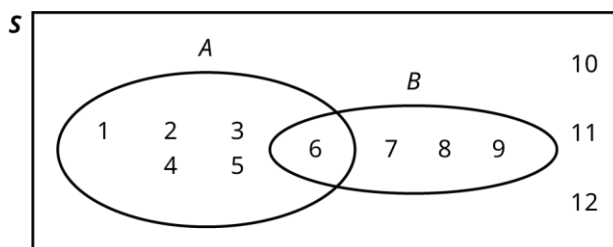
3.5 Sơ đồ Venn

Sơ đồ Venn (Venn diagram) là một hình ảnh minh họa các kết quả của một thực nghiệm. Nó thường bao gồm một hình chữ nhật biểu thị không gian mẫu S cùng với các hình tròn hoặc hình bầu dục. Các hình tròn hoặc hình bầu dục đại diện cho các sự kiện. Sơ đồ Venn cũng giúp chúng ta chuyển đổi các từ ngữ thông thường thành các thuật ngữ toán học, góp phần tăng tính chính xác.

Sơ đồ Venn được đặt tên theo người phát minh ra chúng, John Venn, một giáo sư toán học tại Cambridge và cũng là một mục sư Anh giáo. Công trình chính của ông được thực hiện vào cuối những năm 1870 và đã tạo ra cả một nhánh toán học cũng như một cách tiếp cận mới đối với các vấn đề logic. Chúng ta sẽ phát triển các quy tắc xác suất vừa đề cập bằng cách sử dụng phương pháp này để chứng minh các tiên đề xác suất bao gồm Quy tắc cộng, Quy tắc nhân, Quy tắc phần bù, Tính độc lập, và Xác suất có điều kiện.

Ví dụ 3.27

Giả sử một thực nghiệm có các phương án 1, 2, 3, ..., 12 trong đó mỗi phương án có cơ hội xảy ra như nhau. Gọi sự kiện $A = \{1, 2, 3, 4, 5, 6\}$ và sự kiện $B = \{6, 7, 8, 9\}$. Khi đó $A \cap B = \{6\}$ và $A \cup B = \{1, 2, 3, 4, 5, 6, 7, 8, 9\}$. Sơ đồ Venn như sau:

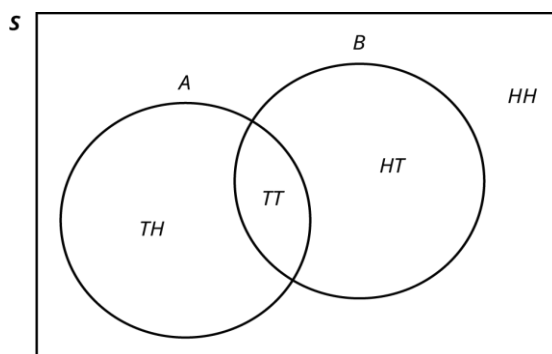
**Hình 3.6****Bài tập vận dụng 3.27**

Giả sử một thực nghiệm có các phương án đen, trắng, đỏ, cam, vàng, xanh lá cây, xanh dương và tím, trong đó mỗi phương án có cơ hội xảy ra như nhau. Gọi sự kiện $C = \{\text{xanh lá cây, xanh dương, tím}\}$ và sự kiện $P = \{\text{đỏ, vàng, xanh dương}\}$. Khi đó $C \cap P = \{\text{xanh dương}\}$ và $C \cup P = \{\text{xanh lá cây, xanh dương, tím, đỏ, vàng}\}$. Vẽ sơ đồ Venn biểu diễn tình huống này.

Ví dụ 3.28

Tung hai đồng xu công bằng. Gọi A = mặt úp ở đồng xu đầu tiên. Gọi B = mặt úp ở đồng xu thứ hai. Khi đó $A = \{TT, TH\}$ và $B = \{TT, HT\}$. Do đó, $A \text{ VÀ } B = \{TT\}$. $A \text{ HOẶC } B = \{TH, TT, HT\}$.

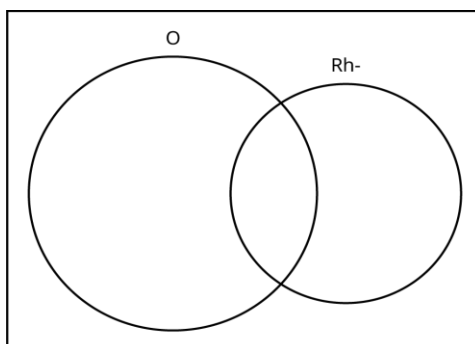
Không gian mẫu khi bạn tung hai đồng xu công bằng là $X = \{HH, HT, TH, TT\}$. Phương án HH KHÔNG nằm trong A CŨNG KHÔNG nằm trong B . Sơ đồ Venn như sau:

**Hình 3.7****Bài tập vận dụng 3.28**

Tung một con xúc xắc 6 mặt công bằng. Gọi A = mặt có số nguyên tố được tung. Gọi B = mặt có số lẻ được tung. Khi đó $A = \{2, 3, 5\}$ và $B = \{1, 3, 5\}$. Do đó, $A \text{ VÀ } B = \{3, 5\}$. $A \text{ HOẶC } B = \{1, 2, 3, 5\}$. Không gian mẫu cho việc tung một con xúc xắc công bằng là $S = \{1, 2, 3, 4, 5, 6\}$. Vẽ sơ đồ Venn biểu diễn tình huống này.

Ví dụ 3.29

Một người có nhóm máu O và yếu tố Rh âm (Rh-) có thể hiến máu cho bất kỳ người nào có bất kỳ nhóm máu nào. 4% người Mỹ gốc Phi có nhóm máu O và yếu tố RH âm, 5-10% người Mỹ gốc Phi có yếu tố Rh-, và 51% có nhóm máu O.

**Hình 3.8**

Hình tròn "O" biểu diễn người Mỹ gốc Phi có nhóm máu O. Hình bầu dục "Rh-" biểu diễn người Mỹ gốc Phi có yếu tố Rh-.

Chúng ta sẽ lấy trung bình của 5% và 10% và sử dụng 7,5% là phần trăm người Mỹ gốc Phi có yếu tố Rh-. Gọi O = người Mỹ gốc Phi có nhóm máu O và R = người Mỹ gốc Phi có yếu tố Rh-.

a. $P(O) =$ _____

b. $P(R) =$ _____

c. $P(O \text{ VÀ } R) =$ _____

d. $P(O \text{ HOẶC } R) =$ _____

e. Trong sơ đồ Venn, mô tả vùng chồng lấn bằng một câu hoàn chỉnh.

f. Trong sơ đồ Venn, mô tả vùng trong hình chữ nhật nhưng bên ngoài cả hình tròn và hình bầu dục bằng một câu hoàn chỉnh.

Lời giải

a. 0,51;

b. 0,075;

c. 0,04;

d. 0,545;

e. Vùng này biểu diễn những người Mỹ gốc Phi có nhóm máu O và yếu tố Rh-.

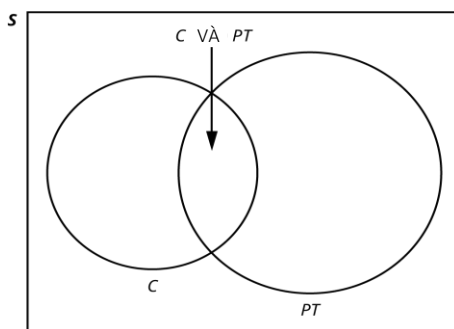
f. Vùng này biểu diễn những người Mỹ gốc Phi không có nhóm máu O cũng không có yếu tố Rh-.

Bài tập vận dụng 3.29

Tung một con xúc xắc 6 mặt công bằng. Gọi A = mặt có số nguyên tố được tung. Gọi B = mặt có số lẻ được tung. Khi đó $A = \{2, 3, 5\}$ và $B = \{1, 3, 5\}$. Do đó, $A \text{ VÀ } B = \{3, 5\}$. $A \text{ HOẶC } B = \{1, 2, 3, 5\}$. Không gian mẫu cho việc tung một con xúc xắc công bằng là $S = \{1, 2, 3, 4, 5, 6\}$. Vẽ sơ đồ Venn biểu diễn tình huống này.

Ví dụ 3.30

40% sinh viên tại một trường đại học địa phương tham gia câu lạc bộ và **50%** làm việc bán thời gian. 5% sinh viên làm việc bán thời gian và tham gia câu lạc bộ. Vẽ sơ đồ Venn thể hiện các mối quan hệ. Gọi C = sinh viên tham gia câu lạc bộ và PT = sinh viên làm việc bán thời gian.



Hình 3.9

Nếu một sinh viên được chọn ngẫu nhiên, tìm

- xác suất sinh viên đó thuộc một câu lạc bộ. $P(C) = 0,40$
- xác suất sinh viên đó làm việc bán thời gian. $P(PT) = 0,50$
- xác suất sinh viên đó thuộc một câu lạc bộ VÀ làm việc bán thời gian. $P(C \text{ VÀ } PT) = 0,05$
- xác suất sinh viên đó thuộc một câu lạc bộ **với điều kiện** sinh viên đó làm việc bán thời gian. $P(C|PT) = \frac{P(C \text{ AND } PT)}{P(PT)} = \frac{0,05}{0,50} = 0,1$
- xác suất sinh viên đó thuộc một câu lạc bộ **HOẶC** làm việc bán thời gian. $P(C \text{ HOẶC } PT) = P(C) + P(PT) - P(C \text{ VÀ } PT) = 0,40 + 0,50 - 0,05 = 0,85$

Để giải Ví dụ 3.30, chúng ta đã phải sử dụng khái niệm xác suất có điều kiện từ phần trước. Ở đó chúng ta đã dùng biểu đồ cây để theo dõi những thay đổi trong xác suất, bởi vì không gian mẫu thay đổi khi chúng ta rút mà không hoàn lại. Nói ngắn gọn, xác suất có điều kiện là khả năng một sự kiện nào đó sẽ xảy ra với điều kiện là một sự kiện khác đã xảy ra trước đó. Nói cách khác, xác suất để một sự kiện nào đó xảy ra với điều kiện là một sự kiện khác cũng đúng. Trong Ví dụ 3.30, xác suất $P(C|PT)$ là xác suất có điều kiện rằng sinh viên được rút ngẫu nhiên là thành viên của câu lạc bộ, với điều kiện là sinh viên đó cũng đang làm việc bán thời gian. Điều này cho phép chúng ta thấy mối quan hệ giữa Sơ đồ Venn và các tiền đề xác suất.

Bài tập vận dụng 3.30

- Năm mươi phần trăm công nhân tại một nhà máy làm công việc thứ hai, 25% có vợ/chồng cũng đi làm, 5% vừa làm công việc thứ hai vừa có vợ/chồng cũng đi làm. Vẽ sơ đồ Venn thể hiện các mối quan hệ này. Gọi W = làm công việc thứ hai và S = vợ/chồng cũng đi làm.
- Tại một hiệu sách, xác suất khách hàng mua tiểu thuyết là 0.6, và xác suất khách hàng mua sách phi hư cấu là 0.4. Giả sử xác suất khách hàng mua cả hai loại là 0.2.
 - Vẽ sơ đồ Venn biểu diễn tình huống này.
 - Tìm xác suất khách hàng mua tiểu thuyết hoặc sách phi hư cấu.
 - Trong sơ đồ Venn, mô tả vùng chồng lấn bằng một câu hoàn chỉnh.
 - Giả sử một số khách hàng chỉ mua đĩa CD. Vẽ một hình bầu dục trong sơ đồ Venn của bạn để biểu diễn sự kiện này.

Ví dụ 3.31

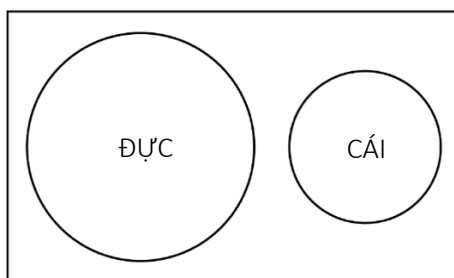
Quan sát một tập hợp gồm 20 con chó Becgie Đức. 12 con đực, 8 con cái, 10 con có một số màu nâu, và 5 con có một số phần lông trắng. Trả lời các câu sau bằng Sơ đồ Venn.

Vấn đề

- Vẽ sơ đồ Venn đơn giản thể hiện tập hợp chó đực và chó cái.
- Vẽ sơ đồ Venn thứ hai minh họa rằng 10 con chó đực có màu nâu.
- Bây giờ vẽ tình huống mô tả một trường hợp trong đó vùng không được tô biểu thị "Không có lông trắng và cái," hoặc $Lông\ trắng' \cap C\acute{a}i$. Dấu phẩy phía trên "lông" chỉ "không có lông trắng." Dấu phẩy phía trên một tập hợp có nghĩa là không thuộc tập hợp đó, ví dụ A' có nghĩa là không phải A . Đôi khi, ký hiệu được sử dụng là một đường kẻ phía trên chữ cái. Ví dụ, $\bar{A} = A'$.

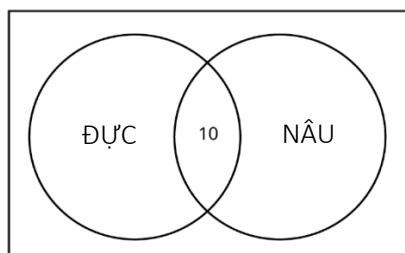
Lời giải

- Sơ đồ Venn dưới đây minh họa tình huống của các sự kiện xung khắc khi các kết quả là các sự kiện độc lập. Một con chó không thể vừa đực vừa cái, do đó không có giao nhau. Là đực thì loại trừ việc là cái và là cái thì loại trừ việc là đực: trong trường hợp này, đặc tính giới tính do đó là xung khắc. Sơ đồ Venn thể hiện điều này như hai tập hợp không có giao nhau. Giao nhau được gọi là tập rỗng sử dụng ký hiệu toán học $\emptyset$.



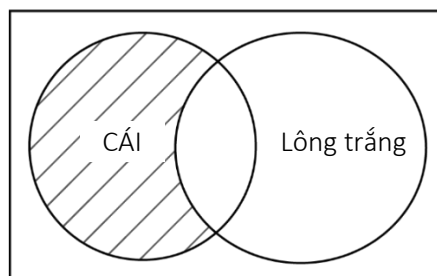
Hình 3.10

- Sơ đồ Venn dưới đây thể hiện sự giao nhau giữa đực và nâu với số 10 được đặt trong đó. Điều này đại diện cho $\text{Đực} \cap \text{Nâu}$: vừa đực vừa nâu. Đây là giao của hai đặc tính này. Để có hợp của Đực và Nâu, thì đơn giản là hai vùng tròn trừ đi phần giao nhau. Theo thuật ngữ chính xác, $\text{Đực} \cup \text{Nâu} = \text{Đực} + \text{Nâu} - \text{Đực} \cap \text{Nâu}$ sẽ cho chúng ta số con chó trong hợp của hai tập hợp này. Nếu chúng ta không trừ giao nhau, chúng ta sẽ đếm lặp một số con chó.



Hình 3.11

c.



Hình 3.12

Quy tắc cộng xác suất

Chúng ta đã gặp quy tắc cộng trước đó nhưng không có sự trợ giúp của sơ đồ Venn. Sơ đồ Venn giúp hình dung việc đếm số lượng vượt quá trong phép tính xác suất. Quy tắc cộng xác suất:

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

Hãy nhớ rằng xác suất chỉ đơn giản là tỷ lệ của các đối tượng mà chúng ta quan tâm so với tổng số đối tượng. Đây là lý do tại sao chúng ta có thể sử dụng tính hữu ích của sơ đồ Venn. Ví dụ 3.31 cho thấy cách chúng ta có thể sử dụng sơ đồ Venn để đếm số chó trong hợp của nâu và đực bằng cách nhắc nhở chúng ta trừ phần chồng lấn của nâu và đực. Chúng ta có thể thấy tác động của điều này trực tiếp đối với xác suất trong quy tắc cộng.

Bài tập vận dụng

Một tập hợp 40 người được chọn ngẫu nhiên. 24 là nam, 16 là nữ, 21 thích cà phê, và 19 thích trà. Trả lời các câu hỏi sau bằng sơ đồ Venn.

- Vẽ sơ đồ Venn đơn giản thể hiện tập hợp nam và nữ.
- Vẽ sơ đồ Venn thứ hai minh họa 14 nam thích cà phê.
- Vẽ tình huống với vùng mô tả nữ và không thích trà.

Ví dụ 3.32

Hãy lấy mẫu 50 sinh viên đang học lớp thống kê. 20 là sinh viên năm nhất và 30 là sinh viên năm hai. 15 sinh viên được điểm “B” trong khóa học, và 5 sinh viên vừa được điểm “B” vừa là sinh viên năm nhất.

Vấn đề

Tìm xác suất chọn được một sinh viên vừa được điểm “B” HOẶC là sinh viên năm nhất. Chúng ta đang chuyển từ HOẶC thành ký hiệu toán học cho quy tắc cộng, đó là hợp của hai tập hợp.

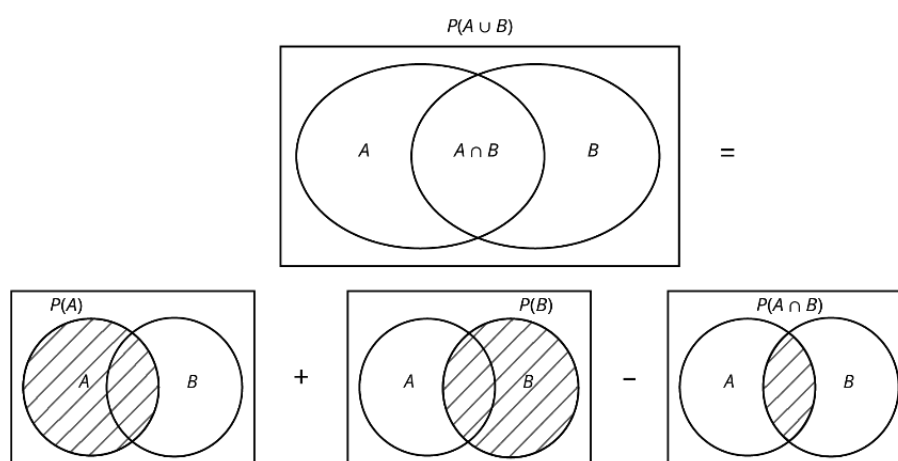
Lời giải

Chúng ta biết rằng có 50 sinh viên trong mẫu của chúng ta, vì vậy chúng ta biết mẫu số của phân số để cho chúng ta xác suất. Chúng ta chỉ cần tìm số sinh viên đáp ứng các đặc tính mà chúng ta quan tâm, tức là bất kỳ sinh viên năm nhất nào và bất kỳ sinh viên nào được điểm “B”. Với Quy tắc Cộng Xác suất, chúng ta có thể bỏ qua trực tiếp việc chuyển sang xác suất.

Gọi “A” = số sinh viên năm nhất, và gọi “B” = điểm “B”. Dưới đây chúng ta có thể thấy quy trình sử dụng Sơ đồ Venn để gán nhãn cho điều này.

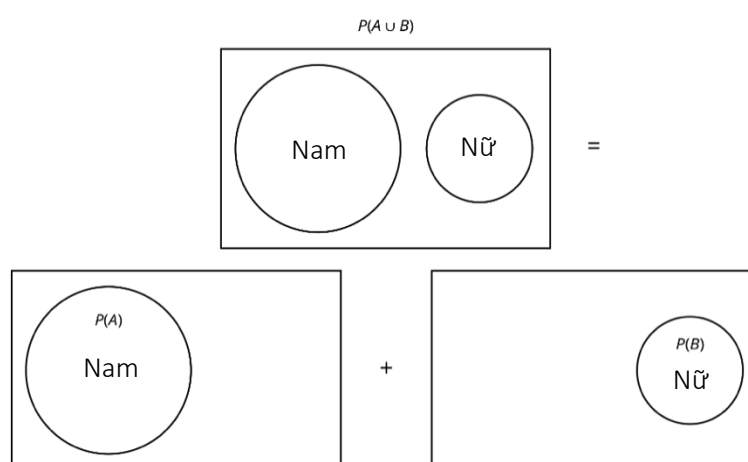
Khi đó, $P(A) = 20/50 = 0.40$, $P(B) = 15/50 = 0.30$, và $P(A \cap B) = 5/50 = 0.10$.

Do đó, $P(A \cup B) = 0.40 + 0.30 - 0.10 = 0.60$.



Hình 3.13

Nếu hai sự kiện xung khắc lẫn nhau, thì, giống như ví dụ khi chúng ta vẽ sơ đồ chó đực và chó cái, quy tắc cộng được đơn giản hóa thành chỉ $P(A \cup B) = P(A) + P(B) - 0$. Điều này đúng bởi vì, như chúng ta đã thấy trước đó, hợp của các sự kiện xung khắc lẫn nhau là tập rỗng, $\emptyset$. Các sơ đồ dưới đây minh họa điều này.



Hình 3.14

Bài tập vận dụng 3.32

Một tập hợp 40 người ngẫu nhiên được chọn. 15 là nam, và 25 là nữ. 18 người thích cà phê. 10 người là nam và thích cà phê. Tìm xác suất chọn được một người thích cà phê hoặc là nam.

Quy tắc nhân xác suất

Nhắc lại Quy tắc nhân xác suất (Multiplication Rule of Probability) bằng cách sử dụng ký hiệu của sơ đồ Venn, chúng ta có:

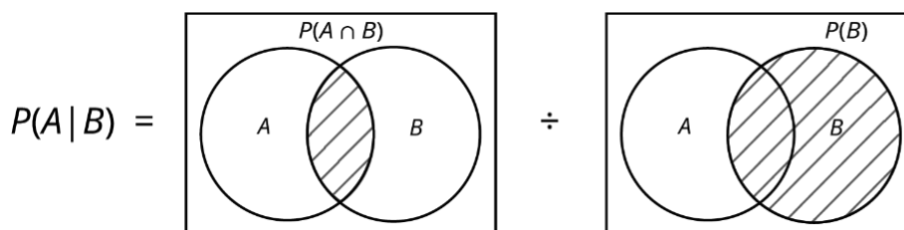
$$P(A \cap B) = P(A|B) \cdot P(B)$$

Quy tắc nhân có thể được sửa đổi với một chút đại số thành quy tắc có điều kiện sau. Sau đó sơ đồ Venn có thể được sử dụng để minh họa quy trình.

Quy tắc có điều kiện: $P(A|B) = P(A \cap B) / P(B)$

Sử dụng các sự kiện tương tự từ Ví dụ 3.32 ở trên, tìm xác suất ai đó sẽ được điểm “B” nếu họ là “sinh viên năm nhất.”

$$P(A|B) = \frac{0,1}{0,3} = \frac{1}{3}$$



Hình 3.15

Quy tắc nhân cũng phải được thay đổi nếu hai sự kiện độc lập. **Các sự kiện độc lập** được định nghĩa là tình huống trong đó xác suất có điều kiện đơn giản là xác suất của sự kiện quan tâm. Chính thức, tính độc lập của các sự kiện được định nghĩa là $P(A|B) = P(A)$ hoặc $P(B|A) = P(B)$. Khi tung đồng xu, kết quả của lần thứ hai độc lập với kết quả của lần đầu tiên; đồng xu không “ghi nhớ” được kết quả của những lần tung trước đó. Quy tắc Nhân Xác suất cho các sự kiện độc lập do đó trở thành:

$$P(A \cap B) = P(A) \cdot P(B)$$

Một cách dễ dàng để nhớ điều này là xem xét điều chúng ta có ý nghĩa bằng từ “và.” Chúng ta thấy rằng Quy tắc Nhân đã chuyển từ “và” thành ký hiệu Venn cho giao. Do đó, kết quả phải đáp ứng hai điều kiện của lần lật đầu tiên và lần thứ hai hoặc một số điều kiện khác. Logic của Quy tắc Nhân Xác suất có thể hiểu được qua việc các phân số khi nhân với nhau sẽ cho ra kết quả nhỏ hơn.

Các Quy tắc Xác suất kết hợp với Sơ đồ Venn sẽ rất hữu ích khi chúng ta cần tính toán xác suất từ dữ liệu trong các bảng phân tích.

Ví dụ 3.33

Bảng 3.13 từ một mẫu 200 người được hỏi về mức độ giáo dục họ đã hoàn thành. Các cột đại diện cho giáo dục cao nhất họ đã hoàn thành, và các hàng phân tách các cá nhân theo nam và nữ.

Bảng 3.13

	Chưa tốt nghiệp THPT	Tốt nghiệp THPT	Chưa tốt nghiệp ĐH	Tốt nghiệp ĐH	Tổng
Nam	5	15	40	60	120
Nữ	8	12	30	30	80
Tổng	13	27	70	90	200

Bây giờ, chúng ta có thể sử dụng bảng này để trả lời các câu hỏi xác suất. Các ví dụ sau được thiết kế để giúp hiểu định dạng ở trên trong khi kết nối kiến thức với cả Sơ đồ Venn và các quy tắc xác suất.

Vấn đề

- Xác suất để một người được chọn vừa tốt nghiệp cao đẳng vừa là phụ nữ là bao nhiêu?
- Xác suất chọn được một phụ nữ hoặc ai đó đã tốt nghiệp cao đẳng là bao nhiêu?
- Xác suất chọn được một người tốt nghiệp trung học nếu chúng ta chỉ chọn từ nhóm nam là bao nhiêu?
- Chúng ta có thể kết luận rằng mức độ giáo dục đạt được bởi 200 người này độc lập với giới tính của người đó không?

Lời giải

- Đây là nhiệm vụ đơn giản của việc tìm giá trị tại vị trí hai đặc tính giao nhau trên bảng, và sau đó áp dụng tiên đề xác suất, nêu rằng xác suất của một sự kiện là tỷ lệ các kết quả phù hợp với sự kiện mà chúng ta quan tâm như một tỷ lệ của tất cả các kết quả có thể.

$$P(\text{Tốt nghiệp Cao đẳng} \cap \text{Nữ}) = 30/200 = 0,15$$

- Câu hỏi này liên quan đến việc sử dụng quy tắc cộng để giải xác suất này.

$$P(\text{Tốt nghiệp Cao đẳng} \cup \text{Nữ}) = P(F) + P(CG) - P(F \cap CG)$$

$$P(\text{Tốt nghiệp Cao đẳng} \cup \text{Nữ}) = \frac{80}{200} + \frac{90}{200} - \frac{30}{200} = \frac{140}{200} = 0,70$$

- Ở đây chúng ta phải sử dụng quy tắc xác suất có điều kiện (quy tắc nhân đã sửa đổi) để giải xác suất này.

$$P(\text{HS Grad} | \text{Man}) = \frac{P(\text{HS Grad} \cap \text{Man})}{P(\text{Man})} = \frac{\frac{15}{200}}{\frac{120}{200}} = \frac{15}{120} = 0,125$$

- Có hai cách để tiếp cận bài kiểm tra này.

Phương pháp đầu tiên tìm cách kiểm tra xem giao của hai sự kiện có bằng tích của các sự kiện riêng biệt hay không, nhớ rằng nếu hai sự kiện độc lập thì $P(A) \cdot P(B) = P(A \cap B)$. Để đơn giản, chúng ta có thể sử dụng các giá trị đã tính từ trên.

$$P(\text{Tốt nghiệp Cao đẳng} \cap \text{Phụ nữ}) = P(\text{CG}) \cdot P(\text{F})?$$

$$\frac{30}{200} \neq \frac{90}{200} \cdot \frac{80}{200} \text{ bởi vì } 0,15 \neq 0,18.$$

Do đó, giới tính và giáo dục ở đây **không** độc lập.

Phương pháp thứ hai là kiểm tra xem xác suất có điều kiện của A cho B có bằng xác suất của A hay không. Một lần nữa để đơn giản, chúng ta có thể sử dụng giá trị đã tính từ trên.

$$P(\text{HS Grad} | \text{Man}) = P(\text{HS Grad})?$$

$$\frac{15}{120} \neq \frac{27}{200} \text{ bởi vì } 0,125 \neq 0,135.$$

Do đó, một lần nữa giới tính và giáo dục ở đây **không** độc lập

Bài tập vận dụng 3.33

Bảng thể hiện sở thích trái cây của 100 người được chọn ngẫu nhiên. Các cột biểu thị loại trái cây, các hàng chia theo giới tính.

Bảng 3.14

	Nho	Dưa hấu	Táo	Lựu	Tổng
Nam	12	18	20	10	60
Nữ	13	7	15	5	40
Tổng	25	25	35	15	100

Trả lời các câu hỏi sau:

- Xác suất một người được chọn thích lựu và là nam là bao nhiêu?
- Xác suất chọn được một nam hoặc ai đó thích táo là bao nhiêu?
- Xác suất chọn được một người thích dưa hấu nếu chúng ta chọn từ nhóm phụ nữ là bao nhiêu?

Thuật ngữ chính

Bảng tiếp liên/ bảng tương quan (Contingency Table) phương pháp hiển thị phân bố tần số dưới dạng bảng với các hàng và cột để chỉ ra cách hai biến có thể phụ thuộc (tương quan) lẫn nhau; bảng cung cấp cách dễ dàng để tính xác suất có điều kiện.

Các sự kiện/biến cố xung khắc (Mutually Exclusive Events) - Hai sự kiện xung khắc lẫn nhau nếu xác suất cả hai xảy ra cùng lúc bằng không. Nếu các sự kiện A và B xung khắc lẫn nhau, thì $P(A \cap B) = 0$.

Đồng khả năng (Equally Likely) - Mỗi kết quả của một thí nghiệm có cùng xác suất.

Hợp $\cup$ (Union) Một kết quả thuộc sự kiện $A \cup B$ nếu kết quả đó thuộc A hoặc thuộc B hoặc thuộc cả A và B .

Giao $\cap$ (Intersection) Một kết quả thuộc sự kiện $A \cap B$ nếu kết quả đó thuộc cả A và B cùng lúc.

Kết quả/ Phương án (Outcome) một kết quả cụ thể của một thực nghiệm

Không gian mẫu (Sample Space) tập hợp tất cả các kết quả có thể của một thực nghiệm

Sơ đồ cây (Tree Diagram) biểu diễn trực quan hữu ích của không gian mẫu và các sự kiện dưới dạng "cây" với các nhánh được đánh dấu bằng các kết quả có thể cùng với các xác suất liên quan (tần số, tần số tương đối)

Sơ đồ Venn (Venn Diagram) biểu diễn trực quan của không gian mẫu và các sự kiện dưới dạng hình tròn hoặc hình oval thể hiện các giao của chúng

Sự kiện/biến cố (Event) - Một tập con của tập hợp tất cả các kết quả của một thực nghiệm; tập hợp tất cả các kết quả của một thực nghiệm được gọi là **không gian mẫu** và thường được ký hiệu bởi S . Một sự kiện là một tập con tùy ý trong S . Nó có thể chứa một kết quả, hai kết quả, không có kết quả (tập con rỗng), toàn bộ không gian mẫu, và tương tự. Ký hiệu chuẩn cho các sự kiện là các chữ cái viết hoa như A , B , C , và

Sự kiện/biến cố độc lập (Independent Events) - Sự xuất hiện của một sự kiện không có ảnh hưởng đến xác suất xuất hiện của sự kiện khác. Các sự kiện A và B độc lập nếu một trong những điều sau đây đúng:

- $P(A|B) = P(A)$
- $P(B|A) = P(B)$
- $P(A \text{ VÀ } B) = P(A)P(B)$

Thực nghiệm/thí nghiệm (Experiment) - Một hoạt động được lên kế hoạch được thực hiện dưới các điều kiện được kiểm soát

Xác suất có điều kiện (Conditional Probability): khả năng một sự kiện sẽ xảy ra cho biết một sự kiện khác đã xảy ra

Xác suất có điều kiện của A CHO BIẾT B (Conditional Probability of A GIVEN B): $P(A|B)$ là xác suất sự kiện A sẽ xảy ra cho biết sự kiện B đã xảy ra.

Xác suất có điều kiện của một sự kiện cho biết sự kiện khác (Conditional Probability of One Event Given Another Event) : $P(A|B)$ là xác suất sự kiện A sẽ xảy ra cho biết sự kiện B đã xảy ra.

Xác suất - một số từ 0 đến 1, bao gồm cả hai đầu, cho biết khả năng một sự kiện cụ thể sẽ xảy ra; nền tảng của thống kê được đưa ra bởi 3 tiên đề sau (theo A.N. Kolmogorov, 1930s): Gọi S ký hiệu không gian mẫu và A và B là hai sự kiện trong S . Khi đó:

- $0 \leq P(A) \leq 1$
- Nếu A và B là hai sự kiện xung khắc lẫn nhau bất kỳ, thì $P(A \cup B) = P(A) + P(B)$.
- $P(S) = 1$

Tóm tắt chương

3.1 Thuật ngữ

Trong mô-đun này chúng ta đã học thuật ngữ cơ bản của xác suất. Tập hợp tất cả các kết quả có thể của một thí nghiệm được gọi là không gian mẫu. Các sự kiện là tập con của không gian mẫu, và chúng được gán một xác suất là một số từ 0 đến 1, bao gồm cả hai đầu.

3.2 Các sự kiện độc lập và xung khắc lẫn nhau

Hai sự kiện A và B độc lập nếu việc biết một sự kiện đã xảy ra không ảnh hưởng đến khả năng sự kiện kia xảy ra. Nếu hai sự kiện không độc lập, thì chúng ta nói rằng chúng phụ thuộc.

Trong lấy mẫu có hoàn lại, mỗi thành viên của tổng thể được thay thế sau khi được chọn, do đó thành viên đó có khả năng được chọn nhiều hơn một lần, và các sự kiện được coi là độc lập. Trong lấy mẫu không hoàn lại, mỗi thành viên của tổng thể chỉ có thể được chọn một lần, và các sự kiện được coi là không độc lập. Khi các sự kiện không có chung kết quả nào, chúng xung khắc lẫn nhau.

3.3 Hai quy tắc cơ bản của xác suất

Quy tắc nhân và quy tắc cộng được sử dụng để tính xác suất của A và B cũng như xác suất của A hoặc B cho hai sự kiện cho trước A, B được định nghĩa trên không gian mẫu. Trong lấy mẫu có hoàn lại, mỗi thành viên của tổng thể được thay thế sau khi được chọn, do đó thành viên đó có khả năng được chọn nhiều hơn một lần, và các sự kiện được coi là độc lập. Trong lấy mẫu không hoàn lại, mỗi thành viên của tổng thể chỉ có thể được chọn một lần, và các sự kiện được coi là không độc lập. Các sự kiện A và B xung khắc lẫn nhau khi chúng không có bất kỳ kết quả chung nào.

3.4 Bảng tiếp liên và sơ đồ cây

Có một số công cụ chúng ta có thể sử dụng để giúp tổ chức và sắp xếp dữ liệu khi tính xác suất. Bảng tiếp liên giúp hiển thị dữ liệu và đặc biệt hữu ích khi tính các xác suất có nhiều biến phụ thuộc.

Sơ đồ cây sử dụng các nhánh để thể hiện các kết quả khác nhau của thí nghiệm và giúp các câu hỏi xác suất phức tạp dễ hình dung.

3.5 Sơ đồ Venn

Sơ đồ Venn là hình ảnh biểu diễn các kết quả của một thực nghiệm. Nó thường bao gồm một hình chữ nhật biểu thị không gian mẫu S hoặc vũ trụ của các đối tượng quan tâm cùng với các hình tròn hoặc hình bầu dục. Các hình tròn hoặc hình bầu dục biểu thị các nhóm sự kiện được gọi là tập hợp. Sơ đồ Venn đặc biệt hữu ích để hình dung một sự kiện, giao của sự kiện, phần bù của một sự kiện và để hiểu các xác suất có điều kiện. Sơ đồ Venn đặc biệt hữu ích để hình dung Giao của hai sự kiện, Hợp của hai sự kiện, hoặc Phần bù của một sự kiện. Hệ thống sơ đồ Venn cũng có thể giúp hiểu các xác suất có điều kiện. Sơ đồ Venn giúp hình dung các phép tính số học thông qua hình ảnh trực quan. Sơ đồ Venn giúp hình dung các phép tính số học thông qua hình ảnh trực quan.

Tóm tắt công thức

3.1 Thuật ngữ

A và B là các sự kiện

$P(S) = 1$ trong đó S là không gian mẫu

$0 \leq P(A) \leq 1$

$P(A|B) = P(A \cap B)/P(B)$

3.2 Các sự kiện độc lập và xung khắc lẫn nhau

Nếu A và B độc lập, $P(A \cap B) = P(A)P(B)$, $P(A|B) = P(A)$ và $P(B|A) = P(B)$.

Nếu A và B xung khắc lẫn nhau, $P(A \cup B) = P(A) + P(B)$ và $P(A \cap B) = 0$.

3.3 Hai Quy tắc cơ bản của xác suất

Quy tắc nhân: $P(A \cap B) = P(A|B)P(B)$

Quy tắc cộng: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Tài liệu tham khảo

3.1 Thuật ngữ

“Countries List by Continent.” Worldatlas, 2013. Available online at <http://www.worldatlas.com/cntycont.htm> (accessed May 2, 2013).

3.2 Sự kiện độc lập và loại trừ lẫn nhau

Lopez, Shane, Preety Sidhu. “U.S. Teachers Love Their Lives, but Struggle in the Workplace.” Gallup Wellbeing, 2013. <http://www.gallup.com/poll/161516/teachers-love-lives-struggle-workplace.aspx> (accessed May 2, 2013).

Data from Gallup. Available online at www.gallup.com/ (accessed May 2, 2013).

3.3 Hai qui tắc cơ bản của xác suất

DiCamillo, Mark, Mervin Field. “The File Poll.” Field Research Corporation. Available online at <http://www.field.com/fieldpollonline/subscribers/Rls2443.pdf> (accessed May 2, 2013).

Rider, David, “Ford support plummeting, poll suggests,” The Star, September 14, 2011. Available online at http://www.thestar.com/news/gta/2011/09/14/ford_support_plummeting_poll_suggests.html (accessed May 2, 2013).

“Mayor’s Approval Down.” News Release by Forum Research Inc. Available online at http://www.forumresearch.com/forms/NewsArchives/NewsReleases/74209_TO_Issue_s_-_Mayoral_Approval_%28Forum_Research%29%2820130320%29.pdf (accessed May 2, 2013).

“Roulette.” Wikipedia. Available online at <http://en.wikipedia.org/wiki/Roulette> (accessed May 2, 2013).

Shin, Hyon B., Robert A. Kominski. “Language Use in the United States: 2007.” United States Census Bureau. Available online at <http://www.census.gov/hhes/socdemo/language/data/acs/ACS-12.pdf> (accessed May 2, 2013).

Data from the Baseball-Almanac, 2013. Available online at www.baseball-almanac.com (accessed May 2, 2013).

Data from U.S. Census Bureau.

Data from the Wall Street Journal.

Data from The Roper Center: Public Opinion Archives at the University of Connecticut. Available online at <http://www.ropercenter.uconn.edu/> (accessed May 2, 2013).

Data from Field Research Corporation. Available online at www.field.com/fieldpollonline (accessed May 2, 2013).

3.4 Bảng tiếp liên và sơ đồ cây

“Blood Types.” American Red Cross, 2013. Available online at <http://www.redcrossblood.org/learn-about-blood/blood-types> (accessed May 3, 2013).

Data from the National Center for Health Statistics, part of the United States Department of Health and Human Services.

Data from United States Senate. Available online at www.senate.gov (accessed May 2, 2013).

Haiman, Christopher A., Daniel O. Stram, Lynn R. Wilkens, Malcom C. Pike, Laurence N. Kolonel, Brien E. Henderson, and Loïc Le Marchand. “Ethnic and Racial Differences in the Smoking-Related Risk of Lung Cancer.” *The New England Journal of Medicine*, 2013. Available online at <http://www.nejm.org/doi/full/10.1056/NEJMoa033250> (accessed May 2, 2013).

“Human Blood Types.” Unite Blood Services, 2011. Available online at <http://www.unitedbloodservices.org/learnMore.aspx> (accessed May 2, 2013).

Samuel, T. M. “Strange Facts about RH Negative Blood.” eHow Health, 2013. Available online at http://www.ehow.com/facts_5552003_strange-rh-negative-blood.html (accessed May 2, 2013).

“United States: Uniform Crime Report – State Statistics from 1960–2011.” The Disaster Center. Available online at <http://www.disastercenter.com/crime/> (accessed May 2, 2013).

3.5 Sơ đồ Venn

Data from Clara County Public H.D.

Data from the American Cancer Society.

Data from The Data and Story Library, 1996. Available online at <http://lib.stat.cmu.edu/DASL/> (accessed May 2, 2013).

Data from the Federal Highway Administration, part of the United States Department of Transportation.

Data from the United States Census Bureau, part of the United States Department of Commerce.

Data from USA Today.

“Environment.” The World Bank, 2013. Available online at <http://data.worldbank.org/topic/environment> (accessed May 2, 2013).

“Search for Datasets.” Roper Center: Public Opinion Archives, University of Connecticut, 2013. Available online at http://www.ropercenter.uconn.edu/data_access/data/search_for_datasets.html (accessed May 2, 2013).

4 BIẾN NGẪU NHIÊN RỜI RẠC



Hình 4.1 Bạn có thể sử dụng xác suất và các biến ngẫu nhiên rời rạc để tính toán khả năng sét sẽ đánh xuống đất 5 lần trong cơn giông kéo dài suốt nửa giờ. (Nguồn: Hình ảnh được lấy từ tác phẩm “CG sét đánh” của Axel Rouvin/Flickr, CC BY 2.0)

Giới thiệu

Một học sinh làm một bài kiểm tra trắc nghiệm đúng sai gồm 10 câu hỏi. Do quá bận rộn, học sinh đó không thể ôn bài và chọn đáp án ngẫu nhiên cho mỗi câu. Vậy xác suất để học sinh đó đạt ít nhất 70% là bao nhiêu?

Một người quản lý của đại lý ô tô quan tâm đến sở thích màu sắc của khách hàng khi mua xe mới. Giả sử đại lý bán trung bình 20 xe mỗi tháng, vậy xác suất để một khách hàng thích xe màu đỏ là bao nhiêu?

Hai ví dụ trên minh họa hai loại vấn đề xác suất khác nhau liên quan đến biến ngẫu nhiên rời rạc. Hãy nhớ rằng dữ liệu rời rạc là dữ liệu có thể đếm được, tức là biến ngẫu nhiên chỉ có thể nhận các giá trị số nguyên. **Một biến ngẫu nhiên (random variable)** mô tả kết quả của một thực nghiệm thống kê bằng lời. Giá trị của biến ngẫu nhiên có thể thay đổi theo mỗi lần lặp lại của một thực nghiệm, hay còn gọi là phép thử (a trial).

Ký hiệu biến ngẫu nhiên

Chữ cái in hoa như X được dùng để biểu thị biến ngẫu nhiên, trong khi chữ thường như x hoặc y được dùng để biểu thị giá trị cụ thể mà biến đó có thể nhận. **Nếu X là một biến ngẫu nhiên, thì X được mô tả bằng lời, còn x là một con số cụ thể.**

Ví dụ: Giả sử X = số lần xuất hiện mặt ngửa khi tung ba đồng xu công bằng. Không gian mẫu khi tung ba đồng xu là: TTT, THH, HTH, HHT, HTT, THT, TTH, HHH. Khi đó, $x = 0, 1, 2, 3$. Vậy X được mô tả bằng lời, còn x là giá trị rời rạc. Do bạn có thể đếm được các giá trị mà X nhận, và vì kết quả là ngẫu nhiên, nên X là một biến ngẫu nhiên rời rạc.

Hàm phân phối xác suất cho biến ngẫu nhiên¹⁰

Một hàm phân phối xác suất (probability distribution function, PDF) có hai đặc tính chính:

1. Mỗi xác suất đều nằm trong khoảng từ 0 đến 1 (bao gồm cả 0 và 1).
2. Tổng của tất cả xác suất là 1.

Hàm phân phối xác suất là một công thức toán học được sử dụng để tính xác suất xảy ra các loại sự kiện cụ thể trong **một thực nghiệm** (an experiment). Có một điều kỳ diệu là cùng một PDF có thể được sử dụng cho nhiều tình huống khác nhau. Ví dụ: hàm phân phối nhị thức (binomial probability distribution) có thể dùng để tính xác suất khi tung đồng xu, khi trả lời các câu hỏi có/không trong bài kiểm tra, hoặc xác định xu hướng trong các cuộc thăm dò ý kiến – nghĩa là bất kỳ sự kiện nhị phân nào.

Một số hàm phân phối xác suất khác như phân phối hình học, phân phối Poisson, hoặc phân phối mũ được sử dụng để tính xác suất thời gian hỏng hóc của thiết bị, thời gian giữa các lượt khách đến, số cuộc gọi đến trung tâm chăm sóc khách hàng, tốc độ phát triển của vi khuẩn, v.v.

Có một hệ thống các họ phân phối xác suất được sử dụng trong nhiều lĩnh vực như y học, kinh doanh, tài chính, vật lý và kỹ thuật. Tuy nhiên, trong chương này, chúng ta sẽ chỉ tập trung vào một số hàm xác suất tiêu biểu, vì chúng đóng vai trò nền tảng trong thống kê suy luận.

Công thức đếm và công thức tổ hợp

Để nhắc lại, xác suất của một sự kiện A, ký hiệu là $P(A)$, được tính bằng:

$$P(A) = \frac{\text{số phương án xảy ra sự kiện A}}{\text{Tổng số phương án có thể xảy ra trong không gian mẫu}}$$

Khi chúng ta xem xét không gian mẫu cho thực nghiệm tung ba đồng xu công bằng, chúng ta có thể dễ dàng liệt kê toàn bộ không gian mẫu và từ đó đếm được số lượng sự kiện thỏa mãn kết quả mong muốn. Ví dụ nếu X là biến ngẫu nhiên rời rạc biểu diễn số lần xuất hiện mặt ngửa khi tung 3 đồng xu thì ta dễ dàng xác định $x = 0, 1, 2, 3$.

Tuy nhiên, khi số lượng phần tử trong không gian mẫu tăng lên – chẳng hạn như với một bộ bài tây gồm 52 lá – thì việc liệt kê toàn bộ không gian mẫu trở nên bất khả thi. Trong những trường hợp như vậy, chúng ta phải chuyển sang cách tiếp cận khác.

Ta sẽ thấy rằng tính toán xác suất chỉ đơn giản là đếm số sự kiện trong nhóm mà ta quan tâm, rồi chia cho tổng số phần tử trong không gian mẫu. Điều này rất dễ thực hiện nếu chúng ta chỉ cần đếm, ví dụ, số sinh viên năm hai trong một lớp học Thống kê mô tả. Nhưng với các vấn đề phức tạp hơn, có thể rất tốn thời gian để liệt kê từng phương án có thể xảy ra – thậm chí là bất khả thi. Ví dụ, khi tung hai con xúc xắc sáu mặt, không gian mẫu có 36 phương án có thể xảy ra nếu biến ngẫu nhiên là tổng điểm của hai mặt trên của hai con

¹⁰ Hàm phân phối xác suất của biến ngẫu nhiên rời rạc còn được gọi là hàm khối lượng xác suất (probability mass function)

xúc xắc. Khi tung bốn con xúc xắc, số phương án trong không gian mẫu sẽ là $6^4 = 1296$. Còn nếu chia tay bài poker 5 lá từ bộ bài 52 lá, có hơn 2,5 triệu tổ hợp khác nhau. Rõ ràng, việc theo dõi tất cả các khả năng như vậy để tính một xác suất duy nhất sẽ rất tẻ nhạt và phi thực tế.

Vì thế, thay vì liệt kê toàn bộ không gian mẫu, chúng ta sử dụng các công thức đếm – tức là các công thức tổ hợp – để xác định số phần tử trong không gian mẫu và sự kiện quan tâm, từ đó tính xác suất bằng cách chia số lượng tương ứng. Các công thức tổ hợp cho phép chúng ta tính từ một tập hợp cho trước có thể tạo ra được bao nhiêu tập hợp con không phân biệt thứ tự với kích thước xác định. Ở đây, “không phân biệt thứ tự” nghĩa là thứ tự của từng phần tử trong tập không ảnh hưởng đến bản chất của tập hợp. Ví dụ, chia một tay bài gồm {át, át, át, át, vua}, thì các cách sắp xếp khác như {vua, át, át, át, át} hay {át, vua, át, át, át} đều được xem là như nhau, vì số lượng lá bài là như nhau và không quan tâm đến thứ tự xuất hiện.

Công thức tổ hợp:

$$\binom{n}{x} = nC_x = \frac{n!}{x!(n-x)!}$$

Công thức trên cho biết số lượng tập hợp con không phân biệt thứ tự, có kích thước x , có thể được tạo tập có n phần tử. Công thức này được gọi là “ n tổ hợp chập x ” và thường được đọc là “từ n chọn x ”. Ký hiệu dấu chấm than (!) được gọi là giai thừa, và biểu thị phép nhân của tất cả các số tự nhiên từ 1 đến số đó. Ví dụ: $4! = 1.2.3.4 = 24$. Lưu ý, theo quy ước toán học $0! = 1$. Công thức này được gọi là **công thức tổ hợp** (Combinatorial Formula). Nó còn được gọi là **hệ số nhị thức** (binomial coefficient) – thuật ngữ sẽ trở nên rõ ràng hơn khi ta học đến phân phối nhị thức trong các chương tiếp theo. Mặc dù khái niệm tổ hợp đã được biết đến từ xa xưa, Blaise Pascal được công nhận là người có đóng góp lớn trong việc chứng minh chính thức, được công bố vào năm 1653. Ngoài ra, ông còn phát triển một phương pháp tổng quát để tính giá trị tổ hợp, được gọi là Tam giác Pascal. Pascal là một trong những thiên tài nổi bật của thời kỳ Phục hưng khoa học, với các công trình của Galileo, Rene Descartes, Isaac Newton, William Shakespeare, và sự hoàn thiện của phương pháp khoa học – lý do cho các chủ đề thống kê mà chúng ta học hôm nay.

Hãy thử đếm tổng số tổ hợp nếu ta rút hai lá từ bốn lá bài át của bộ bài tây. Đây là một tập hợp không phân biệt thứ tự, vì thứ tự rút bài không quan trọng.

Không gian mẫu có thể được liệt kê như sau:

$$S = \{(\text{Bích, Cờ}), (\text{Bích, Rô}), (\text{Bích, Chuồn}), (\text{Rô, Chuồn}), (\text{Cờ, Rô}), (\text{Cờ, Chuồn}), (\text{Rô, Chuồn})\}$$

Ta thấy có 6 tổ hợp tất cả. Theo cách nói chính thức, ta có sáu tập hợp con không phân biệt thứ tự duy nhất, kích thước 2, có thể được tạo ra từ 4 phần tử duy nhất.

Sử dụng công thức tổ hợp, ta có:

$$\binom{4}{2} = \frac{4!}{(4-2)!2!} = \frac{4 \cdot 3 \cdot 2 \cdot 1}{2 \cdot 1 \cdot 2 \cdot 1} = 6$$

Nếu chúng ta muốn biết số lượng tay bài poker 5 lá duy nhất có thể được tạo ra từ bộ bài tây 52 lá, chúng ta chỉ cần tính theo công thức tổ hợp:

$$\binom{52}{5}$$

Trong đó, 52 là tổng số phần tử duy nhất trong quần thể mà chúng ta đang rút ra, và 5 là kích thước nhóm mà chúng ta đang chọn các phần tử vào.

Với công thức tổ hợp (hoặc hệ số nhị thức), chúng ta có thể đếm số lượng phần tử trong không gian mẫu một cách hiệu quả mà không cần liệt kê từng phần tử. Điều này giúp tiết kiệm rất nhiều thời gian và công sức – bởi vì, như đã nói, việc liệt kê tất cả các tay bài 5 lá từ bộ bài 52 lá sẽ là một nhiệm vụ khổng lồ, kéo dài cả đời. Giờ đây, chúng ta có thể ứng dụng công cụ đếm tổ hợp này vào một trong những hàm phân phối xác suất (PDF) quan trọng nhất trong thống kê rời rạc – đó là phân phối siêu bội (hypergeometric distribution).

Hãy nhớ rằng, hàm phân phối xác suất cho phép chúng ta tính xác suất của các sự kiện ngẫu nhiên rời rạc. Chúng ta chỉ cần thay các giá trị thích hợp vào công thức xác suất và sẽ nhận được xác suất của phương án quan tâm. Tuy nhiên, cần lưu ý rằng các hàm phân phối xác suất chỉ hoạt động chính xác nếu được áp dụng đúng trong phạm vi điều kiện mà chúng được thiết kế cho. Mỗi phân phối – chẳng hạn như nhị thức, siêu bội, hay hình học – đều có các giả định và cấu trúc riêng. Việc áp dụng sai loại phân phối sẽ dẫn đến kết quả xác suất không chính xác, ảnh hưởng đến độ tin cậy của toàn bộ quá trình thống kê suy luận.

LƯU Ý

Sử dụng Excel, R để tìm số tổ hợp

Excel: có thể tìm số tổ hợp với hàm COMBIN(n, k), n : tổng số phần tử, k : số phần tử được chọn.

R: có thể tìm số tổ hợp với hàm choose(n, k), n : tổng số phần tử, k : số phần tử được chọn.

4.1. Phân phối siêu bội

Hàm phân phối xác suất đơn giản nhất trong thống kê rời rạc chính là **phân phối siêu bội** (hypergeometric distribution). Đây là một hàm phân phối cơ bản, vì nó được xây dựng dựa trên sự kết hợp giữa kiến thức xác suất cổ điển như sơ đồ Venn, quy tắc cộng và quy tắc nhân, cùng với công thức tổ hợp.

Trong phân phối siêu bội, chúng ta lấy mẫu không thay thế từ một quần thể hữu hạn và tính xác suất để chọn được x phần tử mong muốn trong n lần lấy mẫu từ quần thể hữu hạn có cỡ N chứa A phần tử ta mong muốn.

Ví dụ, trong bộ bài tây có 52 lá, có 4 lá át, ta muốn biết xác suất được chia một tay bài 5 lá, trong đó có 2 lá át.

Đầu tiên, ta tính số phương án tổ hợp 2 lá át từ 4 lá át sử dụng công thức tổ hợp như sau:

$$\binom{4}{2} = \frac{4!}{(4-2)!2!} = 6$$

Và nếu chúng ta không quan tâm đến những gì chúng ta có trong tay đối với ba lá bài còn lại, chúng ta sẽ tính toán số tổ hợp 3 lá từ 48 lá không át:

$$\binom{48}{3} = \frac{48!}{3!45!} = 17.296$$

Kết hợp tất cả lại, chúng ta có thể tính xác suất để có được chính xác hai con át trong một tay bài poker 5 lá như sau:

$$\frac{\binom{4}{2}\binom{48}{3}}{\binom{52}{5}} = 0,0399$$

Lời giải này thực chất chính là phân phối xác suất được gọi là phân phối siêu bội. Công thức tổng quát của phân phối này là:

$$h(x) = \frac{\binom{A}{x}\binom{N-A}{n-x}}{\binom{N}{n}}$$

Trong đó: x là số phần tử thành công mà chúng ta quan tâm (ví dụ: 2 lá át); N là kích thước của quần thể (bộ bài tây 52 lá); A là số phần tử mà ta quan tâm trong quần thể (ví dụ: 4 lá át trong bộ bài); n là số lần lấy mẫu (5 lần chia bài); $h(x)$ là xác suất để xảy ra x thành công trong n lần chọn, khi có A phần tử quan tâm trong một quần thể gồm N phần tử.

Phân phối siêu bội là một ví dụ điển hình của phân phối xác suất rời rạc, bởi vì không thể có thành công từng phần – nghĩa là, bạn không thể có một ván bài với 2,5 lá át. Nói cách khác, biến ngẫu nhiên rời rạc chỉ nhận giá trị nguyên hoặc số đếm được.

Phân phối xác suất này áp dụng cho những trường hợp mà xác suất thành công thay đổi theo từng lần chọn, tức là các phép thử không độc lập. Nói một cách khác, các sự kiện không độc lập với nhau. Điều này xảy ra là do lấy mẫu không thay thế. Ví dụ, khi rút bài từ một bộ bài, nếu chúng ta không đặt lại lá bài sau khi rút, ta đã lấy mẫu không thay thế, và phân phối siêu bội là hàm phù hợp cho trường hợp này. Tuy nhiên, nếu chúng ta đặt lại mỗi lá bài sau khi rút (tức là lấy mẫu có thay thế), thì phân phối siêu bội sẽ không còn phù hợp, và ta sẽ cần một hàm phân phối xác suất (PDF) khác.

Để có thể áp dụng đúng phân phối siêu bội, cần thỏa mãn ba điều kiện sau:

1. Quần thể phải có thể chia thành hai và chỉ hai tập hợp con độc lập (ví dụ: tập lá bài át và tập lá bài không phải át). Biến ngẫu nhiên X = số phần tử có được từ nhóm quan tâm (ví dụ số lá bài từ tập lá át).
2. Xác suất thành công của phép thử thay đổi sau mỗi phép thử. (Trong ví dụ của chúng ta, xác suất rút được lá át sau mỗi lần chia sẽ thay đổi). Nói cách khác, đây là lấy mẫu không thay thế, và do đó các phép thử không độc lập với nhau.

Biến ngẫu nhiên phải là biến rời rạc, nghĩa là chỉ nhận giá trị nguyên đếm được, chứ không phải biến liên tục. (Ví dụ, không thể có 2,7 con át – kết quả phải là các số nguyên như 0, 1, 2, 3, hoặc 4).

Ví dụ 4.1*Vấn đề*

Một đĩa kẹo chứa 30 viên kẹo dẻo và 20 viên kẹo cao su. Mười viên kẹo được chọn ngẫu nhiên. Xác suất để có 5 trong 10 viên kẹo được chọn là kẹo cao su là bao nhiêu? Hai nhóm là kẹo dẻo và kẹo cao su. Vì câu hỏi xác suất yêu cầu xác suất chọn kẹo cao su, nên nhóm quan tâm (nhóm đầu tiên A trong công thức) là kẹo cao su. Kích thước của nhóm quan tâm (nhóm đầu tiên) là 20. Kích thước của nhóm thứ hai là 30. Kích thước của mẫu là 10 (kẹo dẻo hoặc kẹo cao su). Gọi X = số lượng kẹo cao su trong mẫu 10 viên kẹo. X nhận các giá trị $x = 0, 1, 2, \dots, 10$.

- Biểu đạt xác suất bằng công thức toán học?
- Sử dụng hàm phân phối xác suất siêu bội giải vấn đề?
- Xác suất rút được 5 viên kẹo cao su trong 10 lần rút từ đĩa là bao nhiêu?

Lời giải

- $P(x = 5)$
- $P(x = 5) = \frac{\binom{30}{5}\binom{20}{5}}{\binom{50}{10}}$
- $P(x = 5) = 0,215$

Bài tập vận dụng 4.1

Một túi chứa các miếng gỗ in chữ cái. 44 miếng là nguyên âm và 56 miếng là phụ âm. Bảy miếng được chọn ngẫu nhiên. Bạn muốn biết xác suất để bốn trong bảy miếng là nguyên âm. Nhóm quan tâm là gì, kích thước của nhóm quan tâm là bao nhiêu và kích thước của mẫu là bao nhiêu?

Ví dụ 4.2*Vấn đề*

Các nhân viên TSA tuyên bố rằng họ lựa chọn ngẫu nhiên hành khách để kiểm tra hành lý xách tay bổ sung. Trong một giờ vừa qua, có 100 hành khách đi qua điểm kiểm tra an ninh, trong đó có 5 bà cụ tóc bạc. Trong 20 hành khách được chọn để kiểm tra tăng cường, có 3 người là bà cụ tóc bạc. Hãy tính xác suất tương ứng và đưa ra nhận xét?

Lời giải

Bước 1. Xác định biến ngẫu nhiên

Gọi X là biến ngẫu nhiên rời rạc, biểu thị số lượng bà cụ tóc bạc được chọn trong mẫu kích thước 20. Ta quan tâm đến số lượng phần tử từ nhóm đặc biệt (bà cụ tóc bạc) được chọn trong quá trình lấy mẫu không thay thế. Vì X là số đếm, nên đây là biến ngẫu nhiên rời rạc, và có thể nhận các giá trị $X = 0, 1, 2, 3, 4, 5$.

Bước 2. Xác định thực nghiệm

Thực nghiệm thống kê ở đây là quá trình chọn 20 hành khách để kiểm tra kỹ lưỡng hành lý. Theo mô tả, hành khách được chọn ngẫu nhiên, và mỗi hành khách chỉ được chọn một lần, nên đây là lấy mẫu không thay thế.

Bước 3. Xác định phân phối xác suất phù hợp

Quần thể gồm 100 hành khách được chia thành hai nhóm rạch ròi: Nhóm quan tâm (A): bà cụ tóc bạc, kích thước $A = 5$. Nhóm còn lại: tất cả hành khách khác, kích thước $N - A = 95$. Ta chọn mẫu kích thước $n = 20$ và quan sát số người thuộc nhóm quan tâm. Vì vậy, vấn đề thỏa mãn đầy đủ điều kiện của phân phối siêu bội, nên có thể sử dụng hàm phân phối xác suất siêu bội để giải.

Bước 4. Diễn giải câu hỏi xác suất

Câu hỏi được đặt ra là gì? Không giống như Ví dụ 4.1 về việc rút kẹo từ đĩa, không có câu hỏi rõ ràng nào được đặt ra ở đây. Vấn đề đĩa kẹo mang tính cơ học với một câu hỏi rõ ràng được nêu trực tiếp liên quan đến xác suất rút được 5 viên kẹo cao su. Nói tóm lại, một câu hỏi rõ ràng không có hàm ý. Chúng ta có thực sự quan tâm đến xác suất rút được 5 viên kẹo cao su không? Ở đây có một câu hỏi thú vị vì văn phòng an ninh đưa ra một tuyên bố mà chúng ta có thể gián tiếp kiểm tra bằng xác suất đơn giản. TSA tuyên bố họ chọn ngẫu nhiên hành khách để kiểm tra tăng cường. Từ dữ liệu, chúng ta thấy ngay rằng trong số năm hành khách là bà cụ tóc bạc, ba người đã được chọn để kiểm tra bổ sung. Nghĩa là 60% tổng số bà cụ tóc bạc. Chúng ta có thể tự đặt câu hỏi về xác suất chọn 3 trong 5 hành khách trong loại này nếu chúng ta chọn ngẫu nhiên. Hãy nhớ rằng lựa chọn ngẫu nhiên có nghĩa là mỗi đối tượng có xác suất được chọn như nhau. Câu hỏi xác suất của chúng ta bằng lời là "Xác suất chọn trúng 3 bà cụ tóc bạc từ nhóm 100 hành khách là bao nhiêu nếu nhóm chỉ bao gồm 5 bà cụ tóc bạc?" Một cách hình thức, $P(X = 3)$ theo phân phối xác suất siêu bội là bao nhiêu?

Từ công thức: $X = 3$, $A = 5$, $N = 100$ (tổng số hành khách từ đó được chọn để kiểm tra an ninh), $n = 20$ (số lượng hành khách sẽ được chọn để kiểm tra).

$$P(x) = \frac{\binom{5}{3}\binom{95}{17}}{\binom{100}{20}} = 0,048$$

Kết luận của chúng ta là xác suất của việc có 3 trong số 5 bà ngoại tóc bạc được chọn ngẫu nhiên để kiểm tra tăng cường chỉ là 0,048. Xác suất chỉ lớn hơn một phần nghìn so với việc rút 5 lá bài từ bộ bài 52 lá và có được một đôi át. Nói tóm lại, đây là một xác suất rất nhỏ để xảy ra trong thực tế. Một nhận xét có thể rút ra ở đây là chúng ta chỉ có thể tin tưởng với mức tin cậy dưới 5% rằng hành khách được chọn trên cơ sở hoàn toàn ngẫu nhiên.

Ví dụ này chứng minh khả năng công cụ phân tích xác suất để trả lời các câu hỏi thực tiễn.

Bài tập vận dụng 4.2

Có tổng cộng 200 quả bóng trong một túi. Bao gồm 25 quả bóng màu đỏ. 40 quả bóng được chọn ngẫu nhiên từ túi. 8 quả trong số đó được tìm thấy là màu đỏ. Tìm xác suất cho kết quả.

LƯU Ý

Sử dụng Excel, R để tính phân phối siêu bội

Excel: phân phối siêu bội $h(x)$ có thể tính bằng hàm HYPGEOM.DIST(x, n, A, N, 0), x: số phần tử thành công trong mẫu đã chọn; n: số lần lấy mẫu; A: số phần tử ta quan tâm trong quần thể; N: kích thước quần thể; 0 để tính hàm phân phối xác suất, 1 để tính hàm phân phối tích lũy.

R: phân phối siêu bội $h(x)$ có thể tính bằng hàm dhyper(x, A, N- A, n); với x, A, N, n như của hàm Excel.

4.2. Phân phối nhị thức

Một trong những hàm phân phối xác suất (PDF) có nhiều ứng dụng là **phân phối nhị thức** (binomial probability distribution). Phân phối này được sử dụng để tính xác suất cho quá trình nhị thức (binomial process)¹¹. Một quá trình nhị thức là trường hợp trong đó mỗi phép thử (trial) chỉ có hai kết quả có thể xảy ra – được gọi là thành công hoặc thất bại. Đây là một dạng đặc biệt của phép thử Bernoulli (Bernoulli trial), đặt theo tên nhà toán học Jacob Bernoulli, người đã phát triển và mô tả các thuộc tính lý thuyết của quá trình này một cách đầy đủ.

Phân phối nhị thức được đặt tên dựa trên tính chất của hệ thống nhị phân – nơi mọi giá trị đều được biểu diễn bằng 0 và 1 – tương ứng với thất bại (0) và thành công (1) trong các phép thử thống kê.

Công thức nhị thức

$$b(x) = \binom{n}{x} p^x q^{n-x}$$

Trong đó, $b(x)$ là xác suất để có x thành công xảy ra trong n lần thử, khi xác suất thành công trong bất kỳ một phép thử nào là p. Và $q = (1 - p)$ là xác suất thất bại trong mỗi lần thử.

Giờ đây, chúng ta có thể thấy rõ lý do tại sao công thức tổ hợp còn được gọi là hệ số nhị thức, bởi vì nó xuất hiện trực tiếp trong hàm phân phối xác suất của phân phối nhị thức.

¹¹ Còn gọi là thực nghiệm nhị thức (binomial experiment)

Để hàm xác suất nhị thức hoạt động chính xác, cần thỏa mãn điều kiện rằng xác suất thành công (p) trong mỗi phép thử phải không đổi từ lần thử này sang lần thử khác. Nói cách khác, các phép thử phải độc lập – tức là kết quả của mỗi phép thử không bị ảnh hưởng bởi kết quả của các phép thử trước đó.

Một ví dụ điển hình của quá trình nhị thức là tung đồng xu, vì xác suất mặt ngửa trong mỗi lần tung là như nhau và không phụ thuộc vào kết quả của những lần tung trước đó. (Một chút lưu ý, rằng việc sử dụng p để làm ký hiệu cho tham số hàm phân phối nhị thức là vi phạm quy ước ký hiệu thống kê, theo đó tham số quần thể nên được biểu diễn bằng chữ cái Hy Lạp. Trong một số tài liệu học thuật, ký hiệu θ (theta) được sử dụng thay cho p , và đây được xem là cách biểu diễn chính xác hơn theo quy ước thống kê.)

Giống như một tập dữ liệu, một hàm phân phối xác suất có giá trị bình quân và độ lệch chuẩn mô tả tập dữ liệu. Đối với phân phối nhị thức, chúng được đưa ra bởi các công thức:

$$\mu = np$$

$$\sigma = \sqrt{npq}$$

Lưu ý p là tham số duy nhất trong các công thức này. Vì vậy, phân phối nhị thức được xem là thành viên của họ phân phối xác suất một tham số (one-parameter family of probability distributions). Nghĩa là, chỉ cần biết giá trị của p là đủ để mô tả hoàn toàn đặc tính của phân phối nhị thức.

Trong lý thuyết xác suất, trong một số trường hợp nhất định, một phân phối xác suất có thể được sử dụng để xấp xỉ một phân phối khác. Khi đó, ta nói rằng một phân phối là giới hạn (limiting distribution) của phân phối kia.

Một ví dụ là trường hợp lấy mẫu không thay thế một mẫu kích thước nhỏ từ một quần thể lớn. Về mặt lý thuyết lấy mẫu không thay thế vi phạm điều kiện độc lập của phép thử nhị thức nên ta có hàm phân phối chính xác là hàm siêu bội. Tuy nhiên, nếu kích thước mẫu n nhỏ hơn 10% kích thước quần thể N (tức là $n < 0,1N$), thì phân phối nhị thức có thể được sử dụng như một xấp xỉ hợp lý cho phân phối siêu bội. Lý do là khi mẫu rất nhỏ so với quần thể, thì xác suất thành công thay đổi không đáng kể sau mỗi lần rút. Nếu bạn rút bài từ một xấp 6 bộ bài (tức có 24 con át), thì xác suất rút được một con át trong lần rút đầu tiên gần như không ảnh hưởng đến xác suất rút được át trong lần rút thứ hai – không giống như khi bạn chỉ có 4 con át trong một bộ bài.

Một thực nghiệm nhị thức có bốn đặc điểm chính:

1. Có số lượng phép thử xác định, được biểu thị bằng ký hiệu n . Phép thử là sự lặp lại của thực nghiệm.
2. Biến ngẫu nhiên X (số lần thử thành công) là biến rời rạc.
3. Mỗi phép thử có đúng hai kết quả có thể xảy ra: thành công hoặc thất bại, p là xác suất thành công trong một phép thử bất kỳ, còn $q = 1 - p$ là xác suất thất bại.
4. Các phép thử là độc lập và được thực hiện trong cùng điều kiện. Điều này tương tự lấy mẫu có hoàn lại. Các phép thử là độc lập nghĩa là kết quả của phép thử này không ảnh hưởng đến kết quả của các phép thử khác. Hay nói cách khác, xác suất p và q không thay đổi giữa các lần thử. Ví dụ: Giả sử một sinh viên tên Joe đoán ngẫu nhiên các câu hỏi đúng-sai trong môn thống kê. Nếu xác suất đoán đúng

(thành công) là $p = 0,6$, thì xác suất đoán sai (thất bại) là $q = 0,4$. Với mỗi câu hỏi, giá trị của p và q vẫn giữ nguyên trong suốt quá trình làm bài.

Như vậy, kết quả của một thực nghiệm nhị thức sẽ tuân theo phân phối xác suất nhị thức.

Phân phối này có biến ngẫu nhiên X = số lần thành công thu được trong n lần thử nghiệm độc lập; bình quân $\mu = np$; phương sai $\sigma^2 = npq$; độ lệch chuẩn, $\sigma = \sqrt{npq}$.

Bất kỳ thực nghiệm nào có đặc điểm 3, 4 và $n = 1$ được gọi là phép thử Bernoulli (được tên đặt theo Jacob Bernoulli, người vào cuối những năm 1600 đã nghiên cứu sâu về chúng). Ta có một thực nghiệm nhị thức khi ta đếm số lần thành công trong một hoặc nhiều phép thử Bernoulli.

Ví dụ 4.3

Vấn đề

Giả sử bạn chơi một trò chơi mà bạn chỉ có thể thắng hoặc thua. Xác suất thắng mỗi ván là 55%, và xác suất thua là 45%. Mỗi ván chơi là độc lập. Nếu bạn chơi trò chơi đó 20 lần, hãy viết hàm mô tả xác suất rằng bạn thắng 15 trong số 20 lần. Ở đây, nếu bạn định nghĩa X là số lần thắng, thì X nhận các giá trị 0, 1, 2, 3, ..., 20. Xác suất thành công là $p = 0,55$. Xác suất thất bại là $q = 0,45$. Số lần thử là $n = 20$. Câu hỏi xác suất có thể được phát biểu dưới dạng toán học là $P(x = 15)$.

Lời giải

$P(x = 15)$ phát biểu cẩn thận hơn bằng tiếng Anh sẽ là: “xác suất thắng chính xác 15 lần trong 20 lần thử.” $P(x < 15)$ cũng cần được phát biểu cẩn thận, vì nếu ta muốn tìm $P(x < 15)$, ta cần tính tất cả các giá trị của x từ $P(x = 0, 1, 2, \dots, 14)$.

Tương tự, ta cũng có thể muốn biết xác suất “lớn hơn” một số nào đó, điều này yêu cầu phải tính nhiều giá trị của x .

Bài tập vận dụng 4.3

Một huấn luyện viên đang dạy một con cá heo được giải cứu cách bắt cá sống trước khi thả nó về tự nhiên. Xác suất để cá heo bắt cá thành công là 35% và xác suất để cá heo không thực hiện thành công việc này là 65%. Trong 20 lần thử, bạn muốn tìm xác suất để cá heo thành công 12 lần. Tìm $P(X = 12)$ bằng cách sử dụng PDF nhị thức.

Ví dụ 4.4

Vấn đề

Một đồng xu đã được thay đổi trọng lượng để xác suất tung ra mặt ngửa từ 0,5 thành 0,25. Tung xu 5 lần; mỗi lần tung là độc lập. Xác suất để nhận được nhiều hơn 3 mặt ngửa là bao nhiêu? Gọi X = số mặt ngửa trong 5 lần tung đồng xu công bằng. X nhận các giá trị 0, 1, 2, 3, 4, 5. Vì đồng xu được thay đổi để có kết quả $p = 0,25$, nên q là 0,75. Số lần thử nghiệm là $n = 5$. Trình bày câu hỏi bằng công thức toán học.

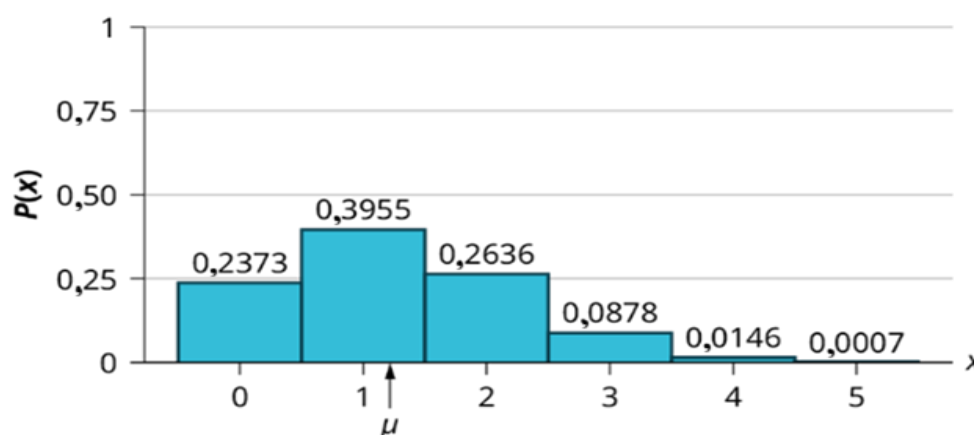
Lời giải

$P(x > 3)$

Đầu tiên, phát triển đầy đủ hàm phân phối xác suất và vẽ đồ thị hàm phân phối xác suất. Với hàm phân phối xác suất được phát triển đầy đủ, chúng ta có thể đơn giản đọc giải pháp cho câu hỏi mặt ngửa $P(x > 3)$.

$P(x > 3) = P(x = 4) + P(x = 5) = 0,0146 + 0,0007 = 0,0153$. Chúng ta đã cộng hai xác suất riêng lẻ theo Quy tắc cộng trong Chương Xác suất.

Hình 4.2 cũng cho phép chúng ta thấy mối liên hệ giữa hàm phân phối xác suất và xác suất và diện tích. Chúng ta cũng thấy trong Hình 4.2 độ lệch của phân phối nhị thức khi p không bằng 0,5. Trong Hình 4.2, phân phối bị lệch phải do $\mu = np = 1,25$ bởi vì $p = 0,25$.



Hình 4.2 Hàm Phân Phối Xác Suất

$$\begin{aligned}
 P(x = x_0) &= \binom{n}{x} p^x (1 - p)^{n-x} \\
 &= \binom{5}{x_0} \cdot 25^{x_0} \cdot 75^{5-x_0} \\
 &\text{etc.} \\
 \mu &= np = 1,25
 \end{aligned}$$

Bài tập vận dụng 4.4

Một con xúc xắc sáu mặt công bằng được tung 10 lần. Mỗi lần tung là độc lập. Bạn muốn tìm xác suất tung được mặt số 1 nhiều hơn ba lần. Trình bày câu hỏi về xác suất bằng công thức toán.

Ví dụ 4.5

Vấn đề

Khoảng 70% sinh viên thống kê làm và nộp bài tập về nhà đúng hạn. Mỗi sinh viên làm bài tập về nhà độc lập. Trong một lớp thống kê gồm 50 sinh viên, xác suất để ít nhất 40 sinh viên làm bài tập về nhà đúng hạn là bao nhiêu? Sinh viên được chọn ngẫu nhiên.

- a. Đây là một vấn đề nhị thức vì chỉ có thành công hoặc _____, có một số lần thử nghiệm cố định và xác suất thành công là 0,70 cho mỗi lần thử nghiệm.
- b. Nếu chúng ta quan tâm đến số lượng sinh viên làm bài tập về nhà đúng hạn, thì chúng ta định nghĩa X như thế nào?
- c. X nhận những giá trị nào?
- d. "Thất bại" bằng lời là gì?
- e. Nếu $p + q = 1$, thì q là gì?
- f. Các từ "ít nhất" được chuyển thành loại bất đẳng thức nào cho câu hỏi xác suất $P(X \geq 40)$.

Lời giải

- a. thất bại.
- b. X = số lượng sinh viên thống kê làm bài tập về nhà đúng hạn.
- c. 0, 1, 2, ..., 50.
- d. Thất bại được định nghĩa là một sinh viên không hoàn thành bài tập về nhà đúng hạn. Xác suất thành công là $p = 0,70$. Số lần thử là $n = 50$.
- e. $q = 0,30$.
- f. lớn hơn hoặc bằng ($\geq$). Câu hỏi xác suất là $P(X \geq 40)$.

Bài tập vận dụng 4.5

1. 65% người vượt qua kỳ thi lái xe cấp bang trong lần thử đầu tiên. Một nhóm gồm 50 cá nhân được chọn ngẫu nhiên từ những người đã tham gia kỳ thi lái xe. Hãy đưa ra hai lý do tại sao đây là một vấn đề nhị thức.
2. Trong một mùa giải NBA, một cầu thủ của Los Angeles Clippers có tỷ lệ hoàn thành cú ném rổ cao nhất trong giải đấu. Cầu thủ này ghi điểm với 61,3% số cú ném của mình. Giả sử bạn chọn một mẫu ngẫu nhiên gồm 80 cú ném được thực hiện bởi cầu thủ này trong mùa giải. Gọi X = số cú ném ghi điểm.
 - a. Phân phối xác suất cho X là gì?
 - b. Sử dụng các công thức, hãy tính (i) giá trị bình quân và (ii) độ lệch chuẩn của X .
 - c. Tìm xác suất để cầu thủ này ghi điểm với 60 cú ném trong số này.
 - d. Tìm xác suất để cầu thủ này ghi điểm với hơn 50 cú ném trong số này.

LƯU Ý

Sử dụng Excel, R để tính phân phối nhị thức.

Excel: phân phối nhị thức $b(x)$ có thể tính bằng hàm BINOM.DIST($x, n, p, 0$); x : số lần thành công mong muốn; n : số lần thử; p : xác suất thành công của mỗi lần thử.

R: phân phối nhị thức $b(x)$ có thể tính bằng hàm `dbinom(x, n, p)`; với x, n, p như của hàm Excel.

4.3. Phân phối hình học

Có bốn đặc điểm chính của một **thực nghiệm hình học** (geometric experiment).

1. Phép thử được lặp lại cho đến khi thành công xảy ra. Hãy nghĩ điều này như thực hiện một hoặc nhiều phép thử Bernoulli với tất cả lần thử thất bại ngoại trừ lần cuối cùng là một thành công. Nói cách khác, bạn tiếp tục lặp lại những gì bạn đang làm cho đến khi đạt được một thành công thì dừng lại. Ví dụ, bạn ném một phi tiêu vào hồng tâm cho đến khi bạn trúng hồng tâm. Ngay khi bạn ném trúng hồng tâm (là một "thành công") thì bạn dừng ném phi tiêu. Có thể mất sáu lần thử cho đến khi bạn trúng hồng tâm: các phép thử như vậy là thất bại, thất bại, thất bại, thất bại, thất bại, thành công. DỪNG LẠI. Về lý thuyết, số lần thử có thể tiếp tục vô tận.
2. Các phép thử lặp lại độc lập với nhau.
3. Xác suất thành công p và xác suất thất bại q là không đổi cho mỗi lần thử. $(p + q) = 1$ và $q = (1 - p)$. Ví dụ, xác suất tung được số 3 khi bạn tung một con xúc xắc công bằng là $\frac{1}{6}$. Điều này đúng bất kể bạn tung xúc xắc bao nhiêu lần. Giả sử bạn muốn biết xác suất nhận được số 3 lần đầu tiên trong lần tung thứ năm. Trong các lần tung từ một đến bốn, bạn không nhận được mặt có số ba. Cho 4 lần tung đầu, xác suất thất bại cho mỗi lần tung là $\frac{5}{6}$. Xác suất nhận được số 3 đầu tiên trong lần tung thứ năm là $(\frac{5}{6})^4 \times \frac{1}{6}$.
4. Biến ngẫu nhiên X biểu thị số lần thử nghiệm mà thành công đầu tiên xảy ra. Tức là, $X =$ số lần thử nghiệm độc lập cho đến thành công đầu tiên.

Các thuộc tính bổ sung sau đây của phân phối hình học:

5. Biến ngẫu nhiên là rời rạc. Biến ngẫu nhiên có thể được định nghĩa theo hai cách tùy thuộc vào mối quan tâm của nhà phân tích. Trong ví dụ trên về việc tung xúc xắc, câu hỏi là "Xác suất thành công đầu tiên sẽ xảy ra trong lần tung thứ năm là bao nhiêu?" Ngoài ra, câu hỏi có thể được hỏi là "Xác suất cần bốn lần thất bại trước khi thành công là bao nhiêu?" Chúng ta sẽ thấy rằng mỗi cách đặt câu hỏi sẽ thay đổi một chút hình thức của hàm phân phối xác suất hình học và sẽ thay đổi giá trị bình quân và độ lệch chuẩn của PDF hình học.
6. Ngầm định trong biến ngẫu nhiên là xác suất thành công là hằng số và do đó xác suất thất bại cũng vậy. Đối với việc tung đồng xu, điều này là hiển nhiên, nhưng trong các thực nghiệm đòi hỏi kỹ năng, chẳng hạn như đánh bóng chày hoặc ném phi tiêu, người ta có thể xem xét rằng việc học trong quá trình thực nghiệm sẽ làm thay đổi xác suất thành công. Phân phối hình học không thể nắm bắt "học tập" do đó lịch sử của xác suất thành công được giả định là không đổi.
7. Phân phối hình học là "không nhớ" (memoryless). Có rất ít hàm phân phối xác suất có đặc tính "không nhớ" và phân phối hình học là hàm duy nhất cho biến ngẫu

nhiên rời rạc là không nhớ. Ví dụ, cầu thủ bóng chày Jones trong lịch sử có thành tích đánh bóng ghi điểm thành công với xác suất 0,20. Jones đã không có cú đánh thành công nào trong suốt 10 lần đánh bóng cuối cùng của mình. Xác suất để Jones có cú đánh thành công trong lần đánh bóng tiếp theo là bao nhiêu? Câu trả lời cho câu hỏi này bỏ qua 10 lần thất bại trước đó của anh ấy. Tất cả các sự kiện trước các sự kiện trong thời gian hiện tại đều không liên quan và do đó được coi là "không nhớ". Xác suất đánh bóng thành công của Jones bắt đầu lại như cũ mỗi khi anh ấy đến đánh bóng. Đặc điểm này trình bày bằng hình thức toán học: $P(x = n + k | x \geq k + 1) = P(x = n)$, trong đó k = số lần thất bại trước đó.

Đặc điểm này của phân phối hình học dẫn đến một kết quả kỳ lạ: nếu ta lấy mẫu một bộ phận từ quy trình sản xuất để kiểm tra lỗi các bộ phận, theo phân phối hình học thì việc kiểm tra trở lại từ đầu bất kể kết quả kiểm tra trước đó. Sẽ nói thêm về điều này khi chúng ta chuyển sang hàm phân phối xác suất hàm mũ.

Ví dụ 4.6

Vấn đề

Bạn chơi một trò chơi may rủi mà bạn có thể thắng hoặc thua (không có khả năng nào khác) cho đến khi bạn thua. Xác suất thua của bạn là $p = 0,57$. Xác suất phải chơi năm ván mới thua là bao nhiêu?

Lời giải

Giả sử X = số ván bạn chơi cho đến khi thua (bao gồm ván thua). Như vậy, X nhận các giá trị 1, 2, 3, ... (có thể tiếp tục vô thời hạn). Câu hỏi về xác suất là $P(x = 5)$.

Bài tập vận dụng 4.6

Bạn ném phi tiêu vào bảng cho đến khi trúng vào vùng trung tâm. Xác suất để bạn trúng vùng trung tâm là $p = 0,17$. Bạn muốn tìm xác suất rằng phải mất tám lần ném mới trúng trung tâm.

Ví dụ 4.7

Vấn đề

Một kỹ sư an toàn tin rằng 35% tất cả các tai nạn công nghiệp trong nhà máy là do nhân viên không tuân theo hướng dẫn. Anh ta quyết định xem xét hồ sơ các báo cáo tai nạn (được chọn ngẫu nhiên và được đặt là hồ sơ sau khi đọc) cho đến khi tìm thấy một báo cáo cho thấy một tai nạn do nhân viên không tuân theo hướng dẫn gây ra. Dự kiến trung bình người kỹ sư an toàn sẽ xem xét bao nhiêu báo cáo cho đến khi tìm thấy một báo cáo cho thấy tai nạn do nhân viên không tuân theo hướng dẫn gây ra? Xác suất người kỹ sư phải kiểm tra ít nhất ba báo cáo cho đến khi tìm thấy một báo cáo cho thấy một tai nạn do nhân viên không tuân theo hướng dẫn gây ra là bao nhiêu?

Gọi X = số lượng báo cáo tai nạn mà kỹ sư an toàn phải kiểm tra cho đến khi tìm thấy một báo cáo cho thấy tai nạn do nhân viên không tuân theo hướng dẫn gây ra. X nhận các giá trị 1, 2, 3, Câu hỏi đầu tiên yêu cầu bạn tìm **giá trị kỳ vọng** (expected value) hoặc giá trị bình quân. Câu hỏi thứ hai yêu cầu bạn tìm $P(\geq 3)$. ("Ít nhất" được chuyển thành ký hiệu "lớn hơn hoặc bằng").

Lời giải

a. Trung bình, người kỹ sư an toàn dự kiến sẽ xem xét bao nhiêu báo cáo cho đến khi tìm thấy một báo cáo cho thấy một tai nạn do không tuân theo hướng dẫn gây ra? Câu hỏi này yêu cầu bạn tìm giá trị kỳ vọng hoặc giá trị bình quân. Đây là số lượng lần thử cho đến khi thành công. Công thức cho giá trị bình quân của phân phối hình học này là: $\mu = \frac{1}{p} = \frac{1}{0,35} = 2,857$.

Lưu ý: rằng giá trị bình quân không cần phải là một số nguyên mặc dù biến ngẫu nhiên là rời rạc nên phải là một số đếm.

b. Xác suất người kỹ sư an toàn phải kiểm tra ít nhất ba báo cáo cho đến khi tìm thấy một báo cáo cho thấy một tai nạn do nhân viên không tuân theo hướng dẫn gây ra là bao nhiêu?

Câu hỏi này được trả lời bằng cách đầu tiên định nghĩa biến ngẫu nhiên. Gọi X = số lượng báo cáo tai nạn mà kỹ sư an toàn phải kiểm tra cho đến khi tìm thấy một báo cáo cho thấy một tai nạn do nhân viên không tuân theo hướng dẫn gây ra. X có thể nhận các giá trị 1, 2, 3, 4, ... ∞ . Không giống như phân phối nhị thức với một số lần thử cố định, phân phối hình học có thể có số lượng lần thử thất bại vô hạn trước khi thành công xảy ra.

Câu hỏi thứ hai này yêu cầu bạn tìm $P(x \geq 3)$. ("Ít nhất" được chuyển thành ký hiệu "lớn hơn hoặc bằng".)

Trong phân phối nhị thức, chúng ta đã thấy giải pháp cho các câu hỏi "nhiều hơn" hoặc "ít hơn" sẽ là tính toán từng xác suất riêng lẻ và cộng từ 0 đến 3. Nếu xác suất quan tâm là "lớn hơn hoặc bằng" ($\geq$), các xác suất riêng lẻ được cộng từ 0 đến 3 và sau đó được trừ đi từ một.

$$P(x \geq 3) = 1 - P(x < 3)$$

Đối với phân phối hình học, chúng ta không thể cộng toàn bộ khoảng giá trị của biến ngẫu nhiên vì nó tiến tới vô cực; ta cần một giải pháp thay thế. Giải pháp là một cặp công thức phân phối hình học cho các câu hỏi yêu cầu xác suất "nhiều hơn" và "ít hơn".

Nếu câu hỏi là:

Xác suất cần NHIỀU HƠN n sự kiện để có thành công đầu tiên là bao nhiêu?

$$P(x > n) = (1 - p)^n$$

Xác suất cần ÍT HƠN n sự kiện để thành công là bao nhiêu?

$$P(x < n) = 1 - (1 - p)^n$$

Bài tập vận dụng 4.7

Một giảng viên cảm nhận rằng 15% sinh viên đạt điểm dưới C trong bài kiểm tra cuối kỳ. Cô ấy quyết định kiểm tra ngẫu nhiên các bài thi cuối kỳ (với lấy mẫu có thay thế, tức là sau khi xem xong, bài thi được đặt trở lại chồng bài) và tiếp tục quá trình này cho đến khi cô ấy tìm thấy bài kiểm tra đầu tiên có điểm dưới C. Tìm xác suất để giảng viên phải kiểm tra ít nhất 10 bài thi trước khi gặp bài đầu tiên có điểm dưới C?

Bài tập vận dụng 4.8

Bạn cần tìm một cửa hàng bán loại mực in đặc biệt. Bạn biết rằng trong số các cửa hàng bán mực in có 10% bán loại mực đặc biệt đó. Bạn gọi điện ngẫu nhiên cho từng cửa hàng cho đến khi một cửa hàng có loại mực bạn cần, p và q là gì?

Ký hiệu cho Phân phối Hình học: G = Hàm Phân phối Xác suất Hình học.

Đọc là "X là một biến ngẫu nhiên có phân phối hình học". Tham số là p ; p = xác suất thành công cho mỗi phép thử.

TRƯỜNG HỢP I: Biến ngẫu nhiên X là số lần thử CHO ĐẾN sự kiện thành công đầu tiên.

Nếu xác suất thành công trong mỗi phép thử là p , thì xác suất của mỗi phép thử thất bại là $(1 - p)$. Ta có xác suất để phép thử x (trong số x lần thử nghiệm) là thành công đầu tiên là:

$$P(X = x) = (1 - p)^{x-1}p$$

Với $x = 1, 2, 3, \dots$

Giống như phân phối nhị thức, phân phối hình học có các tham số là bình quân và độ lệch chuẩn. Giá trị kỳ vọng của X, bình quân cho Trường hợp I, là $\mu = \frac{1}{p}$. Điều này cho chúng ta biết số lần thử nghiệm bình quân cho đến khi chúng ta nhận được thành công đầu tiên. X ở đây bao gồm trong lần thử nghiệm dẫn đến thành công. Dạng phân phối hình học trên được sử dụng để mô hình hóa số phép thử cho đến lần thành công đầu tiên. Số lần thử nghiệm được đếm bao gồm luôn lần thử nghiệm thành công: x = tất cả các lần thử nghiệm bao gồm lần thử nghiệm thành công. Điều này có thể thấy trong công thức tính xác suất. Nếu X = số lần thử nghiệm bao gồm thành công, thì chúng ta phải nhân xác suất thất bại $(1 - p)$ theo số lần thất bại, tức là $(x - 1)$. Độ lệch chuẩn của Trường hợp I của phân phối hình học là:

$$\sigma = \sqrt{\left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right)} = \frac{\sqrt{1 - p}}{p}$$

TRƯỜNG HỢP II: Biến ngẫu nhiên X là số lần thất bại TRƯỚC KHI có sự kiện thành công đầu tiên.

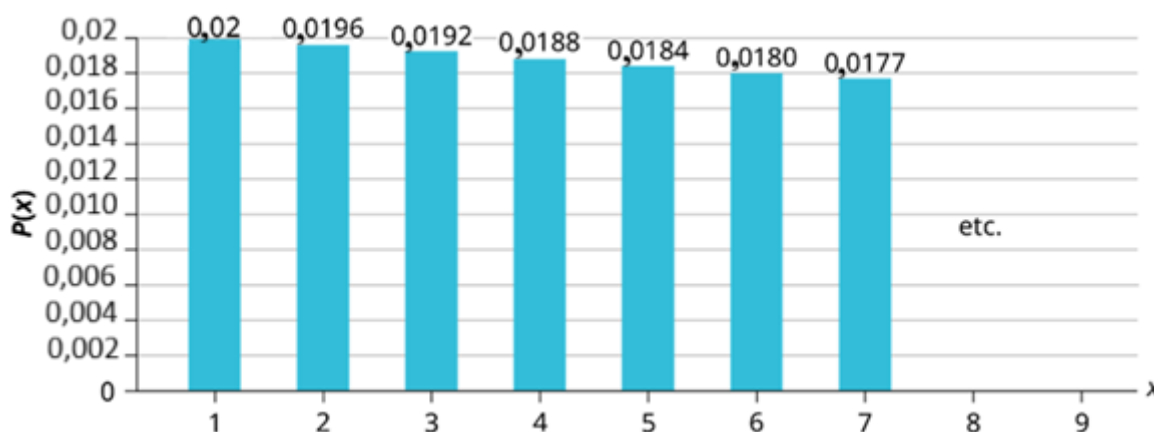
Ngược lại với Trường hợp I, hàm phân phối hình học sau được sử dụng để mô hình hóa số lần thất bại cho đến lần thành công đầu tiên là:

$$P(X = x) = (1 - p)^x p$$

Với $x = 0, 1, 2, 3, \dots$

Trong trường hợp này, lần thử thành công sẽ không được tính công thức: X = số lần thất bại. Giá trị kỳ vọng, bình quân của phân phối này là $\mu = \frac{1-p}{p}$. Điều này cho chúng ta biết có bao nhiêu lần thất bại có thể xảy ra trước khi chúng ta có được thành công. Trong cả hai trường hợp, việc tính toán xác suất đều tuân theo cấp số nhân (geometric sequence). Trong Trường hợp II, độ lệch chuẩn là:

$$\sigma = \sqrt{\left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right)} = \frac{\sqrt{(1-p)}}{p}$$



Hình 4.3

GEOMETRIC: $P = 0,02$, Tỷ số chung $= r = 0,98$

Trục y trong Hình 4.3 chứa xác suất của x . Trục x đại diện cho số lần kiện được kiểm tra cho đến khi phát hiện là linh kiện bị lỗi. Ví dụ, tại $x = 1$, ta có xác suất để linh kiện đầu tiên được tìm thấy bị lỗi là 0,02. Với hai linh kiện được kiểm tra, xác suất linh kiện thứ hai bị lỗi được vẽ tại vị trí $x = 2$ trên trục x với xác suất 0,0196. Đối với xác suất ba lần thử với linh kiện thứ ba bị lỗi, chúng ta tìm thấy $P(X = 3) = 0,019208$.

Ta thấy rằng trong Hình 4.3 rằng các xác suất giảm theo cùng một bước xuống với mỗi thay đổi giá trị của x . Gia số này được gọi là tỷ số chung (common ratio). Điều này chỉ có cho phân phối xác suất hình học. Tỷ số chung, được gọi là r , có thể được tính bằng cách chia giá trị xác suất bất kỳ cho giá trị trước đó, ví dụ: $\frac{P(x=5)}{P(x=4)} = \frac{0,018447}{0,018823} = 0,98$

Đối với tập dữ liệu này, tỷ số chung là 0,98. Tỷ số chung nhân với bất kỳ giá trị xác suất bất kỳ sẽ cung cấp giá trị xác suất tiếp theo trong chuỗi. Ví dụ: xác suất linh kiện thứ sáu được kiểm tra là linh kiện lỗi là 0,018078. Ta có thể kiểm tra điều này bằng công thức từ Trường hợp I. Bây giờ chúng ta có $P(X = 6)$. Biết điều này và tỷ số chung, chúng ta có thể tính toán $P(X = 7)$ bằng phép nhân đơn giản: $P(X = 7) = (0,018078) \times 0,98 = 0,017716$, bằng giá trị chúng ta đã tìm thấy trước đó bằng cách sử dụng phân phối xác suất hình học. Giá số tỷ số chung này là như nhau giữa mọi số và được gọi là cấp số nhân (geometric progression) và đó là lý do mà phân phối xác suất này có tên gọi là geometric distribution. Sau khi tính toán tỷ số chung, bất kỳ giá trị $P(X = x_a)$ nào mong muốn đều có thể dễ dàng tìm thấy.

Số lượng linh kiện bạn dự kiến sẽ kiểm tra cho đến khi tìm thấy linh kiện bị lỗi đầu tiên là giá trị bình quân, $\mu = 50$ cho ví dụ này. Công thức cho bình quân của phân phối hình học cho biến ngẫu nhiên theo định nghĩa $X =$ số lần thất bại trước khi thành công đầu tiên là $\mu = \frac{1}{p} = \frac{1}{0,02} = 50$.

Công thức cho phương sai là

$$\sigma^2 = \left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right) = \left(\frac{1}{0,02}\right)\left(\frac{1}{0,02} - 1\right) = 2.450$$

Độ lệch chuẩn là

$$\sigma = \sqrt{\left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right)} = \sqrt{\left(\frac{1}{0,02}\right)\left(\frac{1}{0,02} - 1\right)} = 49,5$$

Ví dụ 4.9

Vấn đề

Giả sử xác suất một linh kiện máy tính bị lỗi là 0,02. Các linh kiện được chọn ngẫu nhiên. Tìm xác suất linh kiện thứ bảy là linh kiện đầu tiên phát hiện lỗi khi kiểm tra. Bạn dự kiến sẽ kiểm tra bao nhiêu linh kiện cho đến khi tìm thấy một linh kiện bị lỗi?

Lời giải

Gọi $X =$ số lượng linh kiện máy tính được kiểm tra cho đến khi tìm thấy lỗi đầu tiên.

X nhận các giá trị 1, 2, 3, ... trong đó $p = 0,02$.

Tìm $P(X = 7)$. Đáp án: $P(X = 7) = (1 - 0,02)^{(7-1)} \times 0,02 = 0,0177$.

Xác suất kiểm tra tới linh kiện thứ bảy thì phát hiện lỗi đầu tiên là 0,0177.

Đồ thị của $X \sim G(0,02)$ là:

Trục y chứa xác suất của x , trong đó $X =$ số lượng linh kiện máy tính được kiểm tra. Lưu ý rằng các xác suất giảm theo một gia số chung.

Công thức cho phương sai là $\sigma^2 = \left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right) = \left(\frac{1}{0,02}\right)\left(\frac{1}{0,02} - 1\right) = 2.450$

$$\text{Độ lệch chuẩn là } \sigma = \sqrt{\left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right)} = \sqrt{\left(\frac{1}{0,02}\right)\left(\frac{1}{0,02} - 1\right)} = 49,5$$

Bài tập vận dụng 4.9

Xác suất một thanh thép bị lỗi là 0,01. Các thanh thép được chọn ngẫu nhiên. Tìm xác suất lỗi đầu tiên xảy ra trên thanh thép thứ chín.

Ví dụ 4.10

Vấn đề

Nguy cơ mắc ung thư trong suốt cuộc đời là khoảng 1 trên 67 (1,5%). Gọi X = số người bạn hỏi trước khi một người cho biết họ bị ung thư. Biến ngẫu nhiên X trong trường hợp này chỉ bao gồm số lần phép thử thất bại và không tính lần phép thử thành công trong việc tìm một người mắc bệnh. Công thức phù hợp cho biến ngẫu nhiên này là công thức cho Trường hợp II được trình bày ở trên. Khi đó, X là một biến ngẫu nhiên rời rạc với phân phối hình học:

- Xác suất để bạn hỏi 9 người trước khi một người nói họ bị ung thư là bao nhiêu? Điều này tương đương xác suất để bạn hỏi 9 người nói không và người thứ mười nói có là bao nhiêu?
- Xác suất để bạn phải hỏi 20 người là bao nhiêu?
- Tìm (i) giá trị bình quân và (ii) độ lệch chuẩn của X .

Lời giải

a. $P(X = 9) = (1 - 0,015)^9 \times 0,015 = 0,0131$

b. $P(X = 19) = (1 - 0,015)^{19} \times 0,015 = 0,0113$

c. i. Số bình quân $= \mu = \frac{(1-p)}{p} = \frac{(1-0,015)}{0,015} = 65,67$

ii. Độ lệch chuẩn $= \sigma = \sqrt{\frac{(1-p)}{p^2}} = \sqrt{\frac{(1-0,015)}{0,015^2}} = 66,16$

Bài tập vận dụng 4.10

Tỷ lệ biết chữ của một quốc gia được đo lường qua tỷ lệ người từ 15 tuổi trở lên có thể đọc và viết. Tỷ lệ biết chữ của phụ nữ ở Liên hiệp các Thuộc địa Độc lập là 12%. Gọi X = số phụ nữ bạn hỏi trước khi tìm thấy một người nói rằng cô ấy biết chữ.

- Phân phối xác suất của X là gì?

b. Xác suất để bạn hỏi năm phụ nữ trước khi một người nói rằng cô ấy biết chữ là bao nhiêu? c. Xác suất để bạn phải hỏi mười phụ nữ là bao nhiêu?

Ví dụ 4.11

Một cầu thủ bóng chày có tỷ lệ đánh bóng thành công là 0,320. Đây là xác suất chung mà anh ta đánh trúng bóng mỗi khi anh ta ra sân đánh bóng.

Vấn đề

Xác suất để anh ta đánh trúng bóng lần đầu tiên trong lần ra sân đánh bóng thứ ba là bao nhiêu?

Lời giải

$$P(x = 3) = (1 - 0,32)^{3-1} \times 0,32 = 0,1480$$

Trong trường hợp này, trình tự là thất bại, thất bại, thành công.

Vấn đề

Bạn dự kiến cầu thủ này cần bao nhiêu lần ra sân đánh bóng trước khi đánh trúng bóng?

Lời giải

$$\mu = \frac{1}{p} = \frac{1}{0,320} = 3,125 \approx 3$$

Đây đơn giản là giá trị kỳ vọng của X và do đó là giá trị bình quân của phân phối.

Bài tập vận dụng 4.11

Một vận động viên chạy có xác suất về nhất trong một cuộc đua là 0,426.

a. Xác suất họ về nhất trong cuộc đua thứ tư là bao nhiêu?

b. Bạn mong đợi vận động viên cần chạy bao nhiêu cuộc trước khi về nhất?

Ví dụ 4.12

Vấn đề

Có 80% khả năng một con chó Dalmatian có 13 đốm đen. Bạn đến một cuộc thi chó và đếm các đốm trên những con chó Dalmatian. Xác suất để bạn kiểm 3 con chó trước khi tìm thấy một con có 13 đốm đen là bao nhiêu?

Lời giải

$$P(x = 3) = (1 - 0,80)^3 \times 0,80 = 0,0064$$

Bài tập vận dụng 4.12

Có 75% khả năng một con vẹt có màu xanh lá cây. Bạn đến một cửa hàng thú cưng và quan sát những con vẹt. Xác suất bạn sẽ quan sát được 4 con vẹt trước khi tìm thấy một con vẹt có màu xanh lá cây là bao nhiêu?

LƯU Ý

Sử dụng Excel, R để tính phân phối hình học.

Excel: phân phối nhị thức $G(x)$ có thể tính bằng hàm `NEGBINOM.DIST(x, 1, p, 0)`; x : số lần thất bại (thử) trước khi có thành công đầu tiên; p : xác suất thành công của mỗi lần thử.

R: phân phối nhị thức $G(x)$ có thể tính bằng hàm `dgeom(x, p)`; x : số lần thất bại (thử) trước khi có thành công đầu tiên; p : xác suất thành công của mỗi lần thử.

4.4. Phân phối Poisson

Một phân phối xác suất hữu ích khác là **phân phối Poisson** (Poisson Distribution), còn gọi là phân phối thời gian chờ. Phân phối này được sử dụng để xác định cần bao nhiêu nhân viên thu ngân để giữ thời gian chờ xếp hàng ở mức quy định, cần bao nhiêu đường dây điện thoại để tránh hệ thống bị quá tải và nhiều ứng dụng thực tế khác. Một biến thể của phân phối Poisson, phân phối Pascal, được phát minh gần hai thế kỷ trước, ngày nay được các công ty viễn thông trên toàn thế giới sử dụng để tính toán hệ số tải, mức kết nối vệ tinh và các vấn đề về dung lượng Internet. Phân phối này được đặt tên theo Simeon Poisson, người đã trình bày nó vào năm 1837 như một phần mở rộng của phân phối nhị thức mà chúng ta sẽ thấy có thể được ước tính bằng phân phối Poisson.

Có ba đặc điểm chính của một thực nghiệm Poisson.

1. **Phân phối xác suất Poisson** (Poisson Probability Distribution) cho xác suất một số lượng sự kiện xảy ra trong một khoảng thời gian hoặc khoảng không gian xác định nếu các sự kiện này xảy ra với tốc độ trung bình đã biết.
2. Các sự kiện độc lập về mặt thời gian (hoặc không gian) kể từ sự kiện cuối cùng. Ví dụ, một biên tập viên sách có thể quan tâm đến số lượng từ bị đánh sai chính tả trong một cuốn sách cụ thể. Có thể là trung bình có năm từ bị sai chính tả trong 100 trang. Khoảng (không gian) là 100 trang và giả định rằng không có mối quan hệ nào giữa các lỗi chính tả (ví dụ khoảng cách từ một lỗi tới lỗi tiếp theo là ngẫu nhiên).
3. Biến ngẫu nhiên X = số lần xuất hiện trong khoảng thời gian (không gian) quan tâm.

Ví dụ 4.13*Vấn đề*

Một ngân hàng ước tính nhận được trung bình sáu séc xấu mỗi ngày. Xác suất để ngân hàng nhận được ít hơn năm séc xấu trong bất kỳ ngày nào là bao nhiêu? Điều quan tâm là số lượng séc mà ngân hàng nhận được trong một ngày, vì vậy khoảng thời gian quan tâm là một ngày. Gọi X = số lượng séc xấu mà ngân hàng nhận được trong một ngày. Nếu ngân hàng dự kiến nhận được sáu séc xấu mỗi ngày thì giá trị trung bình là sáu séc mỗi ngày. Viết một phát biểu toán học cho câu hỏi xác suất.

Lời giải

$$P(x < 5)$$

Bài tập vận dụng 4.13

Một cửa hàng điện tử dự kiến có trung bình mười lượt trả hàng mỗi ngày. Người quản lý muốn biết xác suất để cửa hàng nhận được ít hơn tám lượt trả hàng trong bất kỳ ngày nào. Phát biểu câu hỏi xác suất bằng toán học.

Ví dụ 4.14*Vấn đề*

Bạn nhận thấy một phóng viên tin tức nói "ờ", trung bình, hai lần mỗi lần phát sóng. Xác suất để phóng viên tin tức nói "ờ" nhiều hơn hai lần mỗi lần phát sóng là bao nhiêu?

Đây là một vấn đề Poisson vì bạn quan tâm đến việc biết số lần phóng viên tin tức nói "ờ" trong một lần phát sóng.

- Khoảng thời gian quan tâm là gì?
- Số lần trung bình phóng viên tin tức nói "ờ" trong một lần phát sóng là bao nhiêu?
- Gọi $X =$ _____. X nhận những giá trị nào?

Câu hỏi xác suất là $P(\text{_____})$.

Lời giải

- một lần phát sóng được đo bằng phút
- 2
- Gọi X = số lần phóng viên tin tức nói "ờ" trong một lần phát sóng. $X = 0, 1, 2, 3, \dots$
- $P(x > 2)$

Bài tập vận dụng 4.14

Phòng cấp cứu tại một bệnh viện tiếp nhận trung bình năm bệnh nhân mỗi giờ. Một bác sĩ muốn biết xác suất phòng cấp cứu tiếp nhận hơn năm bệnh nhân mỗi giờ. Hãy đưa ra lý do tại sao đây là phân phối Poisson.

Ký hiệu cho phân phối Poisson: P = Hàm phân phối xác suất Poisson

$X \sim P(\mu)$

Đọc là "X là một biến ngẫu nhiên có phân phối Poisson." Tham số là μ (hoặc λ); μ (hoặc λ) = giá trị bình quân cho khoảng thời gian chúng ta quan tâm. Giá trị bình quân là số lần xuất hiện xảy ra trung bình trong khoảng thời gian đó.

Công thức để tính xác suất từ một quá trình Poisson là:

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

trong đó $P(x)$ là xác suất của x thành công, μ là số lượng thành công kỳ vọng dựa trên dữ liệu lịch sử, e là logarit tự nhiên xấp xỉ bằng 2,718 và x là số lượng thành công trên một đơn vị, thường là trên một đơn vị thời gian.

Bình quân (số lần xuất hiện trung bình trên một đơn vị thời gian) của phân phối Poisson là μ và độ lệch chuẩn $\sigma = \sqrt{\mu}$.

Để sử dụng phân phối Poisson, một số giả định nhất định phải đúng. Đó là: xác suất của một thành công không thay đổi trong khoảng thời gian xem xét; thành công không thể xảy ra đồng thời trong khoảng thời gian xem xét; và cuối cùng xác suất của một thành công là độc lập (cùng một giả định với phân phối nhị thức).

Ở một số phương diện, phân phối Poisson có thể được coi là một cách thông minh để chuyển đổi một biến ngẫu nhiên liên tục, thường là thời gian, thành một biến ngẫu nhiên rời rạc bằng cách chia thời gian thành các khoảng thời gian độc lập rời rạc. Cách suy nghĩ này về phân phối Poisson giúp chúng ta hiểu tại sao nó có thể được sử dụng để ước tính xác suất cho biến ngẫu nhiên rời rạc từ phân phối nhị thức. Poisson đang hỏi xác suất của một số lượng thành công trong một khoảng thời gian trong khi nhị thức đang hỏi xác suất của một số lượng thành công nhất định cho một số lượng thử nghiệm nhất định.

Ví dụ 4.15

Vấn đề

Leah nhận được khoảng sáu cuộc gọi điện thoại từ 8 giờ sáng đến 10 giờ sáng. Xác suất để Leah nhận được nhiều hơn một cuộc gọi trong 15 phút tiếp theo là bao nhiêu?

Lời giải

Gọi X = số cuộc gọi Leah nhận được trong 15 phút. (Khoảng thời gian quan tâm là 15 phút hoặc $\frac{1}{4}$ giờ.)

$x = 0, 1, 2, 3, \dots$

Nếu Leah nhận được trung bình sáu cuộc gọi điện thoại trong hai giờ và có tám khoảng thời gian 15 phút trong hai giờ, thì Leah nhận được

$\left(\frac{6}{8}\right) = 0,75$ cuộc gọi trong 15 phút, trung bình. Vì vậy, $\mu = 0,75$ cho vấn đề này.

Tìm $P(x > 1)$.

Phân phối Poisson là rời rạc và do đó > 1 bao gồm tất cả các số nguyên đến vô cực. Giải pháp là trừ đi xác suất nhỏ hơn hoặc bằng 1, do đó $1 - [P(x = 0) + P(x = 1)]$

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

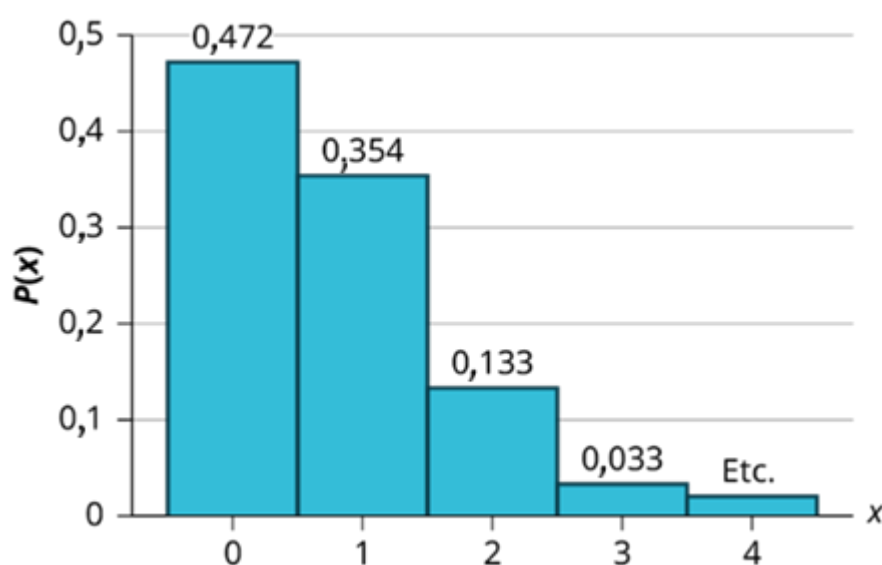
$$P(x > 1) = 1 - P(x \leq 1)$$

$$= 1 - [P(x = 0) + P(x = 1)]$$

$$= 1 - \left[\frac{0,75^0 e^{-0,75}}{0!} + \frac{0,75^1 e^{-0,75}}{1!} \right]$$

$$= 1 - \left[\frac{(1)(0,4724)}{1} + \frac{(0,75)(0,4724)}{1} \right]$$

$$= 1 - [0,4724 + 0,3543] = 0,1733$$



Hình 4.4

Trục y trong Hình 4.4 chứa xác suất của x , trong đó X = số cuộc gọi trong 15 phút.

Bài tập vận dụng 4.15

Một trung tâm dịch vụ khách hàng nhận được khoảng mười email mỗi nửa giờ. Xác suất để trung tâm dịch vụ khách hàng nhận được nhiều hơn bốn email trong sáu phút tiếp theo là bao nhiêu?

Ví dụ 4.16*Vấn đề*

Theo Baydin, một công ty quản lý email, một người dùng email nhận được trung bình 147 email mỗi ngày. Gọi X = số lượng email mà một người dùng nhận được mỗi ngày. Biến ngẫu nhiên rời rạc X nhận các giá trị $x = 0, 1, 2, \dots$. Biến ngẫu nhiên X có phân phối Poisson: $X \sim P(147)$. Giá trị bình quân là 147.

- Xác suất để một người dùng nhận được chính xác 160 email mỗi ngày là bao nhiêu?
- Xác suất để một người dùng nhận được tối đa 160 email mỗi ngày là bao nhiêu?
- Độ lệch chuẩn là bao nhiêu?

Lời giải

- $P(x = 160) = \text{poissonpdf}(147, 160) \approx 0,0180$
- $P(x \leq 160) = \text{poissoncdf}(147, 160) \approx 0,8666$
- Độ lệch chuẩn $= \sigma = \sqrt{\mu} = \sqrt{147} \approx 12,1244$

LƯU Ý

Sử dụng Excel, R để tính phân phối Poisson.

Excel: phân phối Poisson $P(x)$ có thể tính bằng hàm `POISSON.DIST(x, m, 0)`; x : số lần xảy ra sự kiện (giá trị nguyên ≥ 0); m : số lần xảy ra trung bình trong một đơn vị thời gian (hay không gian); có thể trị 0 bằng 1 nếu muốn tính phân phối tích lũy.

R: phân phối nhị thức $P(x)$ có thể tính bằng hàm `dpois(x, m)`; x và m như hàm tính Poisson trong Excel.

Bài tập vận dụng 4.16

Theo một cuộc thăm dò gần đây của Pew Internet Project, những người trong độ tuổi từ 14 đến 17 gửi trung bình 187 tin nhắn mỗi ngày. Gọi X = số lượng tin nhắn mà một cô

gái từ 14 đến 17 tuổi gửi mỗi ngày. Biến ngẫu nhiên rời rạc X nhận các giá trị $x = 0, 1, 2, \dots$. Biến ngẫu nhiên X có phân phối Poisson: $X \sim P(187)$. Giá trị trung bình là 187 tin nhắn văn bản.

- Xác suất để một người gửi chính xác 175 tin nhắn mỗi ngày là bao nhiêu?
- Xác suất để một người gửi tối đa 150 tin nhắn mỗi ngày là bao nhiêu?
- Độ lệch chuẩn là bao nhiêu?

Ví dụ 4.17

Vấn đề

Người dùng tin nhắn nhận hoặc gửi trung bình 41,5 tin nhắn văn bản mỗi ngày.

- Người dùng tin nhắn nhận hoặc gửi bao nhiêu tin nhắn mỗi giờ?
- Xác suất để người dùng tin nhắn nhận hoặc gửi hai tin nhắn mỗi giờ là bao nhiêu?
- Xác suất để người dùng tin nhắn nhận hoặc gửi nhiều hơn hai tin nhắn mỗi giờ là bao nhiêu?

Lời giải

a. Gọi X = số lượng tin nhắn mà người dùng gửi hoặc nhận trong một giờ. Số lượng tin nhắn trung bình nhận được mỗi giờ là $\frac{41,5}{24} \approx 1,7292$.

$$b. P(X = 2) = \frac{\mu^x e^{-\mu}}{x!} = \frac{1,7292^2 e^{-1,7292}}{2!} = 0,265$$

$$c. P(X > 2) = 1 - P(X \leq 2) = 1 - \left[\frac{7^0 e^{-7}}{0!} + \frac{7^1 e^{-7}}{1!} + \frac{7^2 e^{-7}}{2!} \right] = 0,250$$

Bài tập vận dụng 4.17

Sân bay quốc tế Hartsfield-Jackson của Atlanta là sân bay bận rộn nhất thế giới. Trung bình có 2.700 lượt máy bay đến và đi mỗi ngày.

- Có bao nhiêu máy bay đến và đi khỏi sân bay mỗi giờ?
- Xác suất có chính xác 100 lượt đến và đi trong một giờ là bao nhiêu?
- Xác suất có tối đa 100 lượt đến và đi trong một giờ là bao nhiêu?

Ví dụ 4.18

Vấn đề

Vào một ngày cụ thể trong tháng 5, bắt đầu từ 4:30 chiều, xác suất hoạt động địa chấn thấp trong 24 giờ tiếp theo ở Alaska được báo cáo là khoảng 1,02%. Sử dụng thông tin này cho 200 ngày tiếp theo để tìm xác suất có hoạt động địa chấn thấp trong mười trong

số 200 ngày tiếp theo. Sử dụng cả phân phối nhị thức và Poisson để tính xác suất. Chúng có gần nhau không?

Lời giải

Gọi X = số ngày có hoạt động địa chấn thấp.

Sử dụng phân phối nhị thức:

$$P(X = 10) = \frac{200!}{10!(200 - 10)!} \times 0,0102^{10} \times 0,9898^{190} = 0,000039$$

Sử dụng phân phối Poisson:

Tính $\mu = np = 200(0,0102) \approx 2,04$

$$P(X = 10) = \frac{\mu^x e^{-\mu}}{x!} = \frac{2,04^{10} e^{-2,04}}{10!} = 0,000045$$

Khi n lớn (lớn hơn 20) và p nhỏ (nhỏ hơn 0,05) thì kết quả tính xác suất bằng phân phối Poisson sẽ xấp xỉ phân phối nhị thức. Hai kết quả rất gần nhau, cả hai xác suất đều gần bằng 0.

Bài tập vận dụng 4.18

Vào một ngày cụ thể trong tháng 5, bắt đầu từ 4:30 chiều, xác suất hoạt động địa chấn vừa phải trong 24 giờ tiếp theo tại quần đảo Kuril ngoài khơi bờ biển Nhật Bản được báo cáo là khoảng 1,43%. Sử dụng thông tin này cho 100 ngày tiếp theo để tìm xác suất có hoạt động địa chấn thấp trong năm trong số 100 ngày tiếp theo. Sử dụng cả phân phối nhị thức và Poisson để tính xác suất. Chúng có gần giống nhau không?

Ví dụ 4.19

Phân phối Poisson thường được gọi là phân phối cho "thời gian chờ". Theo một nghĩa nào đó, điều này gắn liền với mối liên kết giữa một phân phối rời rạc đo số lần xuất hiện và cách xử lý biến ngẫu nhiên liên tục (thời gian). Phân phối Poisson lấy biến thời gian (ngẫu nhiên liên tục) và chia nó thành một biến ngẫu nhiên rời rạc bằng cách đo số lần xuất hiện rời rạc. Trong ví dụ 4.16, chúng ta đang đo số lượng tin nhắn được gửi (lần xuất hiện), là một biến ngẫu nhiên rời rạc, trong mỗi ngày, là một biến ngẫu nhiên liên tục.

Phân phối Poisson có thể được sử dụng cho nhiều ứng dụng khác. Bất kỳ biến ngẫu nhiên liên tục nào có thể được chia thành các phép đo rời rạc đều có thể sử dụng phân phối Poisson để tính xác suất của số lượng giá trị rời rạc X đã cho, số lần xuất hiện mà chúng ta quan tâm. Chúng ta quan tâm đến số lượng ổ gà trên đường cao tốc để đánh giá chất lượng công việc lát đường. Trong trường hợp này, ổ gà là lần xuất hiện và có thể được đếm trên mỗi dặm vuông đường cao tốc. Dặm là một biến ngẫu nhiên liên tục và ổ gà là biến ngẫu nhiên rời rạc. Có thể chúng ta muốn biết về xác suất mà một kỹ thuật lát đường nhất định sẽ dẫn đến có hơn 20 ổ gà trong 10 dặm mặt đường. Nếu xác suất lớn, chúng ta có thể loại bỏ kỹ thuật lát đường đó. Hay vỏ hộp bánh quy ghi rằng mỗi chiếc bánh

quy có hơn năm miếng sô cô la. Lấy mẫu bột bánh quy, chúng ta có thể tính toán xác suất bánh quy thành phẩm sẽ đáp ứng yêu cầu có năm miếng sô cô la mỗi chiếc bánh quy. Một trường hợp khác mà ta có thể sử dụng phân phối xác suất Poisson. Đó là việc sản xuất kính cho kính chắn gió ô tô có thể gặp phải một lỗi được gọi là "bọt". Trong quá trình sản xuất, không khí đôi khi bị mắc kẹt trong kính nóng và khi làm nguội sẽ để lại một "bọt" ảnh hưởng tới quan sát của người lái xe. Để tính xác suất thành phẩm có ít hơn ba "bọt", số lượng chấp nhận được cho quy trình sản xuất, phân phối Poisson sẽ hữu ích.

LƯU Ý

Poisson là một biến ngẫu nhiên rời rạc. Chúng ta đang đếm số lần xuất hiện (xảy ra). Nó xảy ra hoặc không xảy ra trong một khoảng thời gian, một dặm đường hoặc một mẻ bột bánh quy. Theo nghĩa này, Poisson giống như nhị thức ở chỗ lần xuất hiện là nhị phân, xảy ra hoặc không xảy ra. Vì mối liên kết này với nhị thức, chúng ta có thể sử dụng Poisson để ước tính phân phối nhị thức và điều này được thảo luận dưới đây.

Bài tập vận dụng 4.19

Trong một thành phố nhỏ, một cuộc khảo sát 500 phương tiện cho thấy có 2% khả năng xảy ra tai nạn trong một tuần. Tính xác suất có nhiều nhất 2 vụ tai nạn xảy ra trong bất kỳ tuần nào.

Ước tính Phân phối Nhị thức bằng Phân phối Poisson

Chúng ta đã thấy rằng phân phối nhị thức cung cấp một phép xấp xỉ cho phân phối siêu bội. Bây giờ chúng ta thấy rằng phân phối Poisson có thể cung cấp một phép xấp xỉ cho phân phối nhị thức. Chúng ta nói rằng phân phối nhị thức tiệm cận Poisson. Phân phối nhị thức tiệm cận phân phối Poisson khi n càng lớn và p nhỏ sao cho np trở thành một giá trị không đổi. Có một số quy tắc kinh nghiệm cho việc khi nào ta có thể sử dụng phân phối Poisson để ước tính phân phối nhị thức. Một trong các quy tắc đó là np , giá trị trung bình của nhị thức, phải nhỏ hơn 25. Một số tác giả khác lại cho rằng nó phải nhỏ hơn 7. Có tác giả lưu ý rằng giá trị bình quân và phương sai của phân phối Poisson bằng nhau nên đề xuất rằng np và npq , giá trị bình quân và phương sai của nhị thức, phải lớn hơn 5. Không có một quy tắc kinh nghiệm nào được chấp nhận rộng rãi cho việc khi nào người ta có thể sử dụng Poisson để ước tính phân phối nhị thức.

Khi chúng ta nghiên cứu các phân phối xác suất này, chúng ta đi từ các phân phối ít phức tạp đến các phân phối phức tạp hơn; các phân phối này, có thể nói, chứa các phân phối ít phức tạp hơn trong chúng. Điều này đã được chứng minh bằng toán học. Điều này đưa chúng ta đến mức độ phức tạp cao nhất trong phân phối xác suất tiếp theo, có thể được sử dụng như một phép xấp xỉ cho tất cả những phân phối mà chúng ta đã thảo luận cho đến nay. Đó là phân phối chuẩn.

Ví dụ 4.20*Vấn đề*

Một cuộc khảo sát 500 sinh viên năm cuối của Trường Kinh doanh Price cho thông tin sau. 75% đi làm ngay sau khi tốt nghiệp. 15% tiếp tục làm MBA. 9% ở lại để lấy bằng hai trong một chương trình khác. 1% tiếp tục lấy bằng thạc sĩ tài chính.

Xác suất để có nhiều hơn hai sinh viên tiếp tục học sau đại học để lấy bằng thạc sĩ tài chính là bao nhiêu?

Lời giải

Đây rõ ràng là một vấn đề phân phối xác suất nhị thức. Các lựa chọn là nhị phân khi chúng ta định nghĩa kết quả là "trường sau đại học về tài chính" so với "tất cả các lựa chọn khác". Biến ngẫu nhiên là rời rạc và các sự kiện, chúng ta có thể giả định, là độc lập. Giải quyết như một vấn đề nhị thức, chúng ta có:

Giải pháp Nhị thức

$$n \cdot p = 500 \cdot 0,01 = 5 = \mu$$

$$P(0) = \frac{500!}{0!(500-0)} 0,01^0 (1-0,01)^{500-0} = 0,00675$$

$$P(1) = \frac{500!}{1!(500-1)} 0,01^1 (1-0,01)^{500-1} = 0,03318$$

$$P(2) = \frac{500!}{2!(500-2)} 0,01^2 (1-0,01)^{500-2} = 0,08363$$

$$\text{Cộng tất cả 3 lại} = 0,12339$$

$$1 - 0,12339 = 0,87661$$

Phép tính gần đúng Poisson

$$n \cdot p = 500 \cdot 0,01 = 5 = \mu$$

$$n \cdot p \cdot (1-p) = 500 \cdot 0,01 \cdot (0,99) \approx 5 = \sigma^2 = \mu$$

$$P(X) = \frac{e^{-np}(np)^x}{x!} = \left\{ P(0) = \frac{e^{-5}5^0}{0!} \right\} + \left\{ P(1) = \frac{e^{-5}5^1}{1!} \right\} + \left\{ P(2) = \frac{e^{-5}5^2}{2!} \right\}$$

$$0,0067 + 0,0337 + 0,0842 = 0,1247$$

$$1 - 0,12470 = 0,8753$$

Một phép tính gần đúng sai lệch 1 phần nghìn chắc chắn là một phép tính gần đúng có thể chấp nhận được.

Bài tập vận dụng 4.20

Trong Ví dụ 4.20, xác suất có ít hơn 4 sinh viên năm cuối sẽ theo học chương trình sau đại học để lấy bằng thạc sĩ tài chính là bao nhiêu?

Thuật ngữ chính

Biến ngẫu nhiên (Random Variables - RV) đặc tính của một quần thể đang được nghiên cứu; biến ngẫu nhiên thường được ký hiệu bằng các chữ cái Latinh viết hoa X, Y, Z,...; một giá trị cụ thể của biến từ miền giá trị (tập hợp tất cả các giá trị có thể có của một biến) thường được ký hiệu bằng các chữ cái Latinh viết thường x, y và z. Ví dụ: nếu X là số con trong một gia đình, thì x biểu thị một số nguyên cụ thể 0, 1, 2, 3,... Các biến trong thống kê khác với các biến trong đại số ở hai điểm: (1) Miền giá trị của biến ngẫu nhiên (RV) không nhất thiết phải là một tập hợp số; miền có thể được diễn đạt bằng lời. Ví dụ: nếu X = màu tóc thì miền giá trị là {đen, vàng, xám, xanh lá cây, cam}. (2) Chúng ta chỉ có thể biết giá trị cụ thể x mà biến ngẫu nhiên X nhận sau khi thực hiện thực nghiệm.

Phân phối hình học (Geometric Distribution) Một biến ngẫu nhiên rời rạc phát sinh từ các phép thử Bernoulli; các thử nghiệm được lặp lại cho đến thành công đầu tiên. Biến hình học X được định nghĩa là số lần thử nghiệm cho đến thành công đầu tiên. Giá trị trung bình là $\mu = \frac{1}{p}$ và độ lệch chuẩn là $\sigma = \sqrt{\frac{1}{p} \left(\frac{1}{p} - 1 \right)}$. Xác suất có chính xác x thử nghiệm cho đến thành công đầu tiên được cho bởi công thức: $P(X = x) = p(1 - p)^{(x-1)}$; lần thử nghiệm thứ x là thành công đầu tiên.

Một dạng khác của phân phối hình học là xác suất có x thất bại trước khi có thành công đầu tiên là bao nhiêu? Trong dạng này, lần thử nghiệm có thành công đầu tiên không được tính. Công thức cho cách trình bày này của hình học là: $P(X = x) = p(1 - p)^x$. Giá trị kỳ vọng trong dạng phân phối hình học này là $\mu = \frac{(1-p)}{p}$.

Cách dễ nhất để phân biệt hai dạng phân phối hình học này là ghi nhớ rằng p là xác suất thành công và (1-p) là xác suất thất bại. Trong công thức, các số mũ chỉ đơn giản là đếm số lần thành công và số lần thất bại của kết quả mong muốn của phép thử. Tất nhiên, tổng của hai số này phải cộng lại bằng số lần thử nghiệm trong phép thử.

Phân phối nhị thức (Binomial Distribution) Một biến ngẫu nhiên rời rạc phát sinh từ các phép thử Bernoulli; có một số lượng cố định n các thử nghiệm độc lập. "Độc lập" có nghĩa là kết quả của bất kỳ lần thử nghiệm nào (ví dụ: lần thử nghiệm một) không ảnh hưởng đến kết quả của các lần thử nghiệm tiếp theo và tất cả các lần thử nghiệm được thực hiện trong cùng điều kiện. Trong những trường hợp này, biến ngẫu nhiên nhị thức X được định nghĩa là số lần thành công trong n lần thử nghiệm. Giá trị trung bình là $\mu = np$ và độ lệch chuẩn là $\sigma = \sqrt{npq}$. Xác suất có chính xác x thành công trong n lần thử nghiệm là: $P(X = x) = \binom{n}{x} p^x q^{n-x}$.

Phân phối Poisson (Poisson Distribution) Một biến ngẫu nhiên rời rạc đếm số lần một sự kiện nhất định sẽ xảy ra trong một khoảng thời gian cụ thể. Đặc điểm của biến: (1) Xác suất sự kiện xảy ra trong một khoảng thời gian nhất định là như nhau cho tất cả các khoảng thời gian. (2) Các sự kiện xảy ra với giá trị bình quân đã biết và độc lập với thời gian kể từ sự kiện cuối cùng. Phân phối được định nghĩa bởi giá trị bình quân μ của sự kiện trong khoảng thời gian xác định. Giá trị bình quân $\mu = np$. Độ lệch chuẩn là $\sigma = \sqrt{\mu}$. Xác suất có chính xác x thành công là: $P(x) = \frac{\mu^x e^{-\mu}}{x!}$. Phân phối Poisson thường được sử dụng để xấp xỉ phân phối nhị thức khi n "lớn" và p "nhỏ" (quy tắc chung là n phải lớn hơn hoặc bằng 25 và p phải nhỏ hơn hoặc bằng 0,01).

Phân phối siêu bội (Hypergeometric Distribution) Một biến ngẫu nhiên rời rạc được đặc trưng bởi: (1) Một số lượng cố định các lần thử nghiệm. (2) Xác suất thành công không giống nhau từ lần thử nghiệm này sang lần thử nghiệm khác.

Chúng ta lấy mẫu từ hai nhóm phần tử (kết hợp) trong đó chúng ta chỉ quan tâm đến một nhóm. X được định nghĩa là số lần thành công (lấy được phần tử từ trong nhóm ta quan tâm). Xác suất để có x thành công mẫu ngẫu nhiên (không hoàn lại) n phần tử từ tập N phần tử có A phần tử ta quan tâm là $h(x) = \frac{\binom{A}{x} \binom{N-A}{n-x}}{\binom{N}{n}}$.

(Hàm) phân phối xác suất (Probability Distribution Function - PDF) Một mô tả toán học của một biến ngẫu nhiên rời rạc, được đưa ra dưới dạng một phương trình (công thức) hoặc dưới dạng một bảng liệt kê tất cả các kết quả có thể có của một thực nghiệm và xác suất liên quan đến mỗi kết quả.

Phép thử Bernoulli (Bernoulli trial) Một phép thử với các đặc điểm sau: (1) Chỉ có hai kết quả có thể xảy ra được gọi là "thành công" và "thất bại" cho mỗi lần thử nghiệm. (2) Xác suất p của một thành công là như nhau cho bất kỳ lần thử nghiệm nào (do đó xác suất $q = 1 - p$ của một thất bại là như nhau cho bất kỳ lần thử nghiệm nào).

Thực nghiệm hình học (Geometric Experiment) Một thực nghiệm thống kê với các thuộc tính sau: (1) Có một hoặc nhiều phép thử Bernoulli với tất cả các thất bại ngoại trừ lần cuối cùng, đó là một thành công. (2) Về lý thuyết, số lần thử nghiệm có thể tiếp tục vô tận. Phải có ít nhất một lần thử nghiệm. (3) Xác suất thành công p và xác suất thất bại q không thay đổi từ lần thử nghiệm này sang lần thử nghiệm khác.

Thực nghiệm nhị thức (Binomial Experiment) Một thực nghiệm thống kê đáp ứng ba điều kiện sau: (1) Có một số lần thử nghiệm cố định, n . (2) Chỉ có hai kết quả có thể xảy ra, được gọi là "thành công" và "thất bại", cho mỗi lần thử nghiệm. Chữ p biểu thị xác suất thành công trong một lần thử nghiệm và q biểu thị xác suất thất bại trong một lần thử nghiệm. (3) Số n lần thử nghiệm độc lập và được lặp lại bằng các điều kiện giống hệt nhau.

Thực nghiệm siêu bội (Hypergeometric Experiment) Một thực nghiệm thống kê với các thuộc tính sau: (1) Bạn lấy mẫu từ hai nhóm. (2) Một nhóm trong hai nhóm là nhóm bạn quan tâm, được gọi là nhóm đầu tiên. (3) Bạn lấy mẫu ngẫu nhiên mà không thay thế từ hai nhóm kết hợp. (4) Các lần lấy mẫu (chọn) không độc lập vì lấy mẫu là không thay thế.

Tóm tắt chương

Giới thiệu

Các đặc điểm của một hàm phân phối xác suất (PDF) như sau:

1. Mỗi xác suất nằm trong khoảng từ 0 đến 1, bao gồm cả hai giá trị 0 và 1.
2. Tổng các xác suất là 1.

4.1 Phân phối siêu bội

Công thức tổ hợp có thể cung cấp số lượng tập hợp con duy nhất có kích thước x có thể được tạo từ n phần tử duy nhất để giúp chúng ta tính toán xác suất. Công thức tổ hợp là:

$$\binom{n}{x} = {}_nC_x = C_x^n = \frac{n!}{x!(n-x)!}$$

Một thực nghiệm siêu bội là một thực nghiệm thống kê với các thuộc tính sau:

- Bạn lấy các mẫu từ hai nhóm.
- Bạn quan tâm đến một nhóm cụ thể, gọi là nhóm thứ nhất.
- Bạn lấy mẫu không thay thế từ các nhóm kết hợp.
- Mỗi lần chọn không độc lập với nhau, vì việc lấy mẫu là không thay thế.

Kết quả của một phép thử siêu bội phù hợp với phân phối xác suất siêu bội. Biến ngẫu nhiên X = số lượng phần tử từ nhóm quan tâm $h(x) = \frac{\binom{A}{x}\binom{N-A}{n-x}}{\binom{N}{n}}$.

4.2 Phân phối nhị thức

Một thực nghiệm thống kê có thể được phân loại là một thực nghiệm nhị thức nếu đáp ứng các điều kiện sau:

- Có một số lượng cố định các lần thử nghiệm, n .
- Chỉ có hai kết quả có thể xảy ra, được gọi là "thành công" và "thất bại" cho mỗi lần thử nghiệm. Chữ p biểu thị xác suất thành công trong một lần thử nghiệm và q biểu thị xác suất thất bại trong một lần thử nghiệm.
- n lần thử nghiệm độc lập và được lặp lại bằng các điều kiện giống hệt nhau.

Kết quả của một thực nghiệm nhị thức phù hợp với phân phối xác suất nhị thức. Biến ngẫu nhiên X = số lần thành công thu được trong n lần thử nghiệm độc lập. Giá trị trung bình của X có thể được tính bằng công thức $\mu = np$ và độ lệch chuẩn được cho bởi công thức $\sigma = \sqrt{npq}$.

Công thức cho hàm phân phối xác suất nhị thức là: $P(x) = \frac{n!}{x!(n-x)!} \cdot p^x q^{(n-x)}$

4.3 Phân phối hình học

Có ba đặc điểm của một thực nghiệm hình học:

- Có một hoặc nhiều phép thử Bernoulli với tất cả các thất bại ngoại trừ lần cuối cùng, đó là một thành công.
- Về lý thuyết, số lần thử nghiệm có thể tiếp tục vô tận. Phải có ít nhất một lần thử nghiệm.
- Xác suất p (thành công) và xác suất q (thất bại) là như nhau cho mỗi lần thử nghiệm.

Trong một thực nghiệm hình học, ta định nghĩa biến ngẫu nhiên rời rạc X là số lần thử nghiệm độc lập cho đến thành công đầu tiên. Chúng ta nói rằng X có phân phối hình học và viết $X \sim G(p)$ trong đó p là xác suất thành công trong một lần thử nghiệm.

Số bình quân của phân phối hình học $X \sim G(p)$ là $\mu = \frac{1}{p}$ trong đó x = số lần thử nghiệm cho đến thành công đầu tiên cho công thức $P(X = x) = (1 - p)^{x-1}p$ trong đó số lần thử nghiệm bao gồm lần thử nghiệm thành công đầu tiên.

Một cách phát biểu khác của phân phối hình học đặt ra câu hỏi: xác suất để có x lần thất bại cho đến khi có lần thành công đầu tiên là bao nhiêu? Trong cách phát biểu này, phép thử mang lại thành công đầu tiên sẽ không được đếm. Công thức cho cách trình bày này của phân phối hình học là:

$$P(X = x) = p(1 - p)^x$$

Số bình quân trong dạng phân phối hình học này là $\mu = \frac{1-p}{p}$.

Cách dễ nhất để phân biệt hai dạng này của phân phối hình học là hãy nhớ rằng p là xác suất thành công và $(1 - p)$ là xác suất thất bại. Trong công thức, các số mũ chỉ đơn giản là đếm số lần thành công và số lần thất bại của kết quả mong muốn của phép thử. Tất nhiên, tổng của hai số này phải cộng lại bằng số lần thực hiện các phép thử trong một thử nghiệm.

4.4 Phân phối Poisson

Phân phối xác suất Poisson của một biến ngẫu nhiên rời rạc cho xác suất một số lượng sự kiện xảy ra trong một khoảng thời gian hoặc không gian cố định nếu các sự kiện này xảy ra với tốc độ trung bình đã biết và các sự kiện độc lập về mặt thời gian (hoặc không gian) kể từ sự kiện cuối cùng. Phân phối Poisson có thể được sử dụng để xấp xỉ phân phối nhị thức nếu xác suất thành công là "nhỏ" (nhỏ hơn hoặc bằng 0,01) và số lần thử nghiệm là "lớn" (lớn hơn hoặc bằng 25). Các quy luật kinh nghiệm khác cũng được các tác giả khác nhau đề xuất nhưng tất cả đều đồng ý rằng phân phối Poisson là phân phối giới hạn của phân phối nhị thức khi n tăng và p tiến đến 0.

Công thức để tính xác suất từ một quá trình Poisson là: $P(x) = \frac{\mu^x e^{-\mu}}{x!}$

trong đó $P(X)$ là xác suất thành công, μ là số lượng thành công kỳ vọng, e là logarit tự nhiên xấp xỉ bằng 2,718 và X là số lượng thành công trên một đơn vị, thường là trên một đơn vị thời gian.

Tóm tắt công thức

4.1 Phân phối siêu bội

$$h(x) = \frac{\binom{A}{x} \binom{N-A}{n-x}}{\binom{N}{n}}.$$

4.2 Phân phối nhị thức

$X \sim B(n, p)$ có nghĩa là biến ngẫu nhiên rời rạc X có phân phối xác suất nhị thức với n lần thử nghiệm và xác suất thành công p .

X = số lần thành công trong n lần thử nghiệm độc lập

n = số lần thử nghiệm độc lập

X nhận các giá trị $x = 0, 1, 2, 3, \dots, n$

p = xác suất thành công cho bất kỳ lần thử nghiệm nào

q = xác suất thất bại cho bất kỳ lần thử nghiệm nào

$$p + q = 1$$

$$q = 1 - p$$

Giá trị trung bình của X là $\mu = np$. Độ lệch chuẩn của X là $\sigma = \sqrt{npq}$.

$$P(x) = \frac{n!}{x!(n-x)!} \cdot p^x q^{(n-x)}$$

trong đó $P(X)$ là xác suất có X thành công trong n lần thử nghiệm khi xác suất thành công trong MỘT LẦN THỬ NGHIỆM BẤT KỲ là p .

4.3 Phân phối hình học

$$P(X = x) = p(1-p)^x$$

$X \sim G(p)$ có nghĩa là biến ngẫu nhiên rời rạc X có phân phối xác suất hình học với xác suất thành công trong một lần thử nghiệm p .

X = số lần thử nghiệm độc lập cho đến thành công đầu tiên.

X nhận các giá trị $x = 1, 2, 3, \dots$

p = xác suất thành công cho bất kỳ lần thử nghiệm nào

q = xác suất thất bại cho bất kỳ lần thử nghiệm nào

$$p + q = 1$$

$$q = 1 - p$$

Giá trị trung bình là $\mu = \frac{1}{p}$.

Độ lệch chuẩn là $\sigma = \sqrt{\frac{1-p}{p^2}} = \sqrt{\left(\frac{1}{p}\right)\left(\frac{1}{p} - 1\right)}$

4.4 Phân phối Poisson

$X \sim P(\mu)$ có nghĩa là X có phân phối xác suất Poisson trong đó X = số lần xuất hiện trong khoảng thời gian quan tâm.

X nhận các giá trị $x = 0, 1, 2, 3, \dots$

Giá trị bình quân μ (hoặc λ) thường được cho.

Phương sai là $\sigma^2 = \mu$ và độ lệch chuẩn là $\sqrt{\mu}$.

Nếu $P(\mu)$ được sử dụng để xấp xỉ phân phối nhị thức thì $\mu = np$ trong đó n biểu thị số lần thử nghiệm độc lập và p biểu thị xác suất thành công trong một lần thử nghiệm.

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

Tài liệu tham khảo

4.2 Phân phối nhị thức

“Access to electricity (% of population),” The World Bank, 2013. Available online at http://data.worldbank.org/indicator/EG.ELC.ACCS.ZS?order=wbapi_data_value_2009%20wbapi_data_value%20wbapi_data_value-first&sort=asc (accessed May 15, 2015).

“Distance Education.” Wikipedia. Available online at http://en.wikipedia.org/wiki/Distance_education (accessed May 15, 2013). “NBA Statistics – 2013,” ESPN NBA, 2013. Available online at http://espn.go.com/nba/statistics/_/seasontype/2 (accessed May 15, 2013).

Newport, Frank. “Americans Still Enjoy Saving Rather than Spending: Few demographic differences seen in these views other than by income,” GALLUP® Economy, 2013. Available online at <http://www.gallup.com/poll/162368/americansenjoy-saving-rather-spending.aspx> (accessed May 15, 2013).

Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, Serge Tran. The American Freshman: National Norms Fall 2011. Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also available online at <http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf> (accessed May 15, 2013).

“The World FactBook,” Central Intelligence Agency. Available online at <https://www.cia.gov/library/publications/the-world-factbook/geos/af.html> (accessed May 15, 2013). “What are the key statistics about pancreatic cancer?” American Cancer Society, 2013. Available online at <http://www.cancer.org/cancer/pancreaticcancer/detailedguide/pancreatic-cancer-key-statistics> (accessed May 15, 2013).

4.3 Phân phối hình học

“Millennials: A Portrait of Generation Next,” PewResearchCenter. Available online at <http://www.pewsocialtrends.org/files/2010/10/millennials-confident-connected-open-to-change.pdf> (accessed May 15, 2013). “Millennials: Confident. Connected. Open to Change.” Executive Summary by PewResearch Social & Demographic Trends, 2013. Available online at <http://www.pewsocialtrends.org/2010/02/24/millennials-confident-connected-open-tochange/> (accessed May 15, 2013).

“Prevalence of HIV, total (% of populations ages 15-49),” The World Bank, 2013. Available online at http://data.worldbank.org/indicator/SH.DYN.AIDS.ZS?order=wbapi_data_value_2011+wbapi_data_value+wbapi_data_value-last&sort=desc (accessed May 15, 2013).

Pryor, John H., Linda DeAngelo, Laura Palucki Blake, Sylvia Hurtado, Serge Tran. The American Freshman: National Norms Fall 2011. Los Angeles: Cooperative Institutional Research Program at the Higher Education Research Institute at UCLA, 2011. Also

available online at <http://heri.ucla.edu/PDFs/pubs/TFS/Norms/Monographs/TheAmericanFreshman2011.pdf> (accessed May 15, 2013).

“Summary of the National Risk and Vulnerability Assessment 2007/8: A profile of Afghanistan,” The European Union and ICON-Institute. Available online at http://ec.europa.eu/europeaid/where/asia/documents/afgh_brochure_summary_en.pdf (accessed May 15, 2013).

“The World FactBook,” Central Intelligence Agency. Available online at <https://www.cia.gov/library/publications/the-world-factbook/geos/af.html> (accessed May 15, 2013).

“UNICEF reports on Female Literacy Centers in Afghanistan established to teach women and girls basic resading [sic] and writing skills,” UNICEF Television. Video available online at <http://www.unicefusa.org/assets/video/afghan-femaleliteracy-centers.html> (accessed May 15, 2013).

4.4 Phân Phối Poisson

“ATL Fact Sheet,” Department of Aviation at the Hartsfield-Jackson Atlanta International Airport, 2013. Available online at <http://www.atl.com/about-atl/atl-factsheet/> (accessed February 18, 2019). Center for Disease Control and Prevention. “Teen Drivers: Fact Sheet,” Injury Prevention & Control: Motor Vehicle Safety, October 2, 2012. Available online at http://www.cdc.gov/Motorvehiclesafety/Teen_Drivers/teendrivers_factsheet.html (accessed May 15, 2013).

“Children and Childrearing,” Ministry of Health, Labour, and Welfare. Available online at <http://www.mhlw.go.jp/english/policy/children/children-childrearing/index.html> (accessed May 15, 2013). “Eating Disorder Statistics,” South Carolina Department of Mental Health, 2006. Available online at <http://www.state.sc.us/dmh/anorexia/statistics.htm> (accessed May 15, 2013).

“Giving Birth in Manila: The maternity ward at the Dr Jose Fabella Memorial Hospital in Manila, the busiest in the Philippines, where there is an average of 60 births a day,” theguardian, 2013. Available online at <http://www.theguardian.com/world/gallery/2011/jun/08/philippines-health#/?picture=375471900&index=2> (accessed May 15, 2013).

“How Americans Use Text Messaging,” Pew Internet, 2013. Available online at <http://pewinternet.org/Reports/2011/CellPhone-Texting-2011/Main-Report.aspx> (accessed May 15, 2013).

Lenhart, Amanda. “Teens, Smartphones & Texting: Texting volum is up while the frequency of voice calling is down. About one in four teens say they own smartphones,” Pew Internet, 2012. Available online at http://www.pewinternet.org/~media/Files/Reports/2012/PIP_Teens_Smartphones_and_Texting.pdf (accessed May 15, 2013).

“One born every minute: the maternity unit where mothers are THREE to a bed,” MailOnline. Available online at <http://www.dailymail.co.uk/news/article-2001422/Busiest-maternity-ward-planet-averages-60-babies-day-mothersbed.html> (accessed May 15, 2013).

Vanderkam, Laura. “Stop Checking Your Email, Now.” CNNMoney, 2013. Available online at <http://management.fortune.cnn.com/2012/10/08/stop-checking-your-email-now/> (accessed May 15, 2013).

“World Earthquakes: Live Earthquake News and Highlights,” World Earthquakes, 2012. http://www.worldearthquakes.com/index.php?option=ethq_prediction (accessed May 15, 2013).

5 BIẾN NGẪU NHIÊN LIÊN TỤC



Hình 5.1 Chiều cao của những cây củ cải là các biến ngẫu nhiên liên tục. (nguồn: chỉnh sửa từ tác phẩm “Radish (*Raphanus sativus*): White rust caused by *Albugo candida*” của Scot Nelson/ Flickr, Phạm vi công cộng).

Giới thiệu

Biến ngẫu nhiên liên tục có rất nhiều ứng dụng. Trung bình số điểm ghi trong trận bóng chày, điểm IQ, thời lượng của một cuộc gọi điện đường dài, số tiền một người mang theo, tuổi thọ của một con chip máy tính, suất lợi nhuận từ một khoản đầu tư, và điểm thi SAT chỉ là một vài ví dụ ứng dụng. Việc tính toán độ tin cậy hay các phân tích rủi ro cũng phụ thuộc nhiều vào biến ngẫu nhiên liên tục.

LƯU Ý

Giá trị của biến ngẫu nhiên rời rạc và liên tục đôi khi có thể gây nhầm lẫn. Ví dụ, nếu X là số dặm đường (làm tròn đến dặm gần nhất) bạn lái xe đến chỗ làm, thì X là biến ngẫu nhiên rời rạc. Bạn đếm số dặm. Nếu X là quãng đường bạn lái xe đến nơi làm việc, thì khi đó bạn đo lường các giá trị của X và X sẽ là một biến ngẫu nhiên liên tục. Một ví dụ khác, nếu X là số lượng sách trong ba lô, thì X là biến ngẫu nhiên rời rạc. Nếu X là trọng lượng của một cuốn sách, thì X là biến ngẫu nhiên liên tục vì trọng lượng được đo lường. Định nghĩa biến ngẫu nhiên như thế nào là việc rất quan trọng.

5.1 Các tính chất của hàm mật độ xác suất

Đồ thị của một **phân phối xác suất liên tục** (continuous probability distribution) là một đường cong. Xác suất được biểu diễn bằng diện tích dưới đường cong. Chúng ta đã từng gặp khái niệm này khi xây dựng tần suất tương đối bằng đồ thị histogram trong Chương về Thống kê mô tả. Diện tích tương đối cho một khoảng giá trị chính là xác suất rút ngẫu nhiên một quan sát thuộc nhóm đó (khoảng giá trị). Trong Chương Biến ngẫu nhiên rời rạc, khi mô tả phân phối Poisson, đồ thị trong Ví dụ 4.15 sử dụng các hình hộp để biểu

diễn xác suất của các giá trị cụ thể của biến ngẫu nhiên Poisson. Trong trường hợp này, chúng ta đã “dễ dãi” sử dụng các hình hộp để diễn tả xác suất vì biến ngẫu nhiên ở đây theo phân phối Poisson là biến rời rạc (số nguyên) nên chỉ là một điểm trên trục số; trong khi đó hình hộp lại bề ngang theo trục số. Nếu chú tâm, ta thấy rằng trên đồ thị tại Ví dụ 4.15, mỗi điểm trên trục hoành của đồ thị xác suất gắn với một hộp, tương ứng với một xác suất của giá trị của X . Ngược lại, trên đồ thị histogram, một hộp sẽ không gắn với một điểm. Xác suất để biến ngẫu nhiên liên tục nhận một giá trị cụ thể là bằng 0 vì diện tích dưới một điểm là bằng 0. Xác suất được đo lường bằng diện tích bên dưới đường cong.

Đường cong này được gọi là **hàm mật độ xác suất** (probability density function, viết tắt là **pdf**). Ta sử dụng ký hiệu $f(x)$ để biểu diễn đường cong. $f(x)$ là hàm mô tả đồ thị; ta sử dụng hàm mật độ $f(x)$ để vẽ đồ thị phân phối xác suất.

Diện tích dưới đường cong (area under the curve) được tính bởi một hàm khác gọi là **hàm phân phối tích lũy** (cumulative distribution function, viết tắt là **cdf**). Hàm phân phối tích lũy được sử dụng để tính xác suất như là diện tích. Về mặt toán học, xác suất tích lũy là tích phân của hàm mật độ xác suất, và xác suất cho một khoảng giá trị của biến ngẫu nhiên liên tục sẽ bằng tích phân của pdf trong khoảng đó: tức là diện tích dưới đường cong giữa hai giá trị.

Hãy nhớ rằng diện tích dưới pdf cho tất cả các giá trị khả dĩ của biến ngẫu nhiên luôn bằng 1.

Vì vậy, xác suất có thể được hiểu như phần trăm chắc chắn rằng biến ngẫu nhiên liên tục có giá trị nằm giữa hai giá trị được quan tâm.

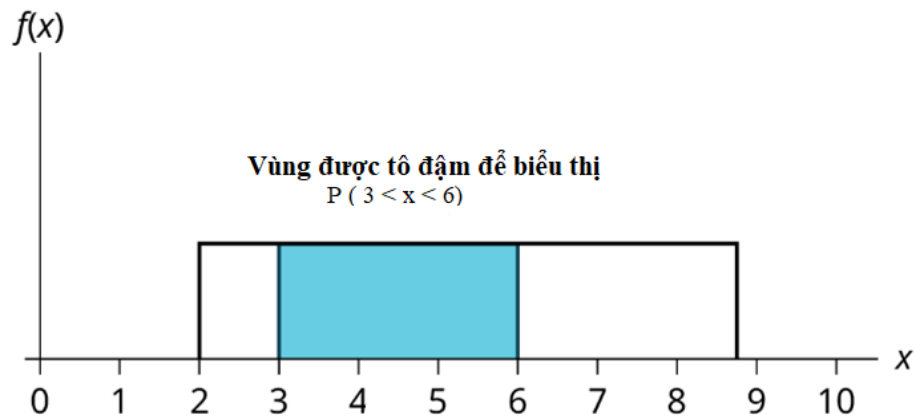
Tóm lại các điểm chính:

- Kết quả được đo lường, không phải đếm.
- Tổng diện tích của phần nằm dưới đường cong và phía trên trục x bằng 1.
- Xác suất được tìm cho các khoảng giá trị của x , không phải cho giá trị riêng lẻ.
- $P(c < x < d)$ là xác suất rằng biến ngẫu nhiên x nằm trong khoảng từ c đến d . $P(c < x < d)$ chính là diện tích dưới đường cong, phía trên trục x , bên phải của c và bên trái của d .
- $P(x = c) = 0$. Xác suất để x nhận một giá trị cụ thể là bằng 0 vì diện tích dưới đường cong từ $x = c$ đến $x = c$ không có bề rộng, nên diện tích = 0.
- $P(c < x < d) = P(c \leq x \leq d)$ vì xác suất là diện tích (không phụ thuộc việc lấy hoặc không lấy ranh giới).

Chúng ta sẽ tìm phần diện tích đại diện cho xác suất bằng cách sử dụng hình học, công thức tính, phần hỗ trợ, hoặc bảng xác suất. Nói chung, cần sử dụng tích phân để tìm diện tích dưới đường cong cho các hàm mật độ xác suất. Các công thức được dùng để tìm diện tích trong tài liệu này đều được suy ra bằng kỹ thuật tích phân.

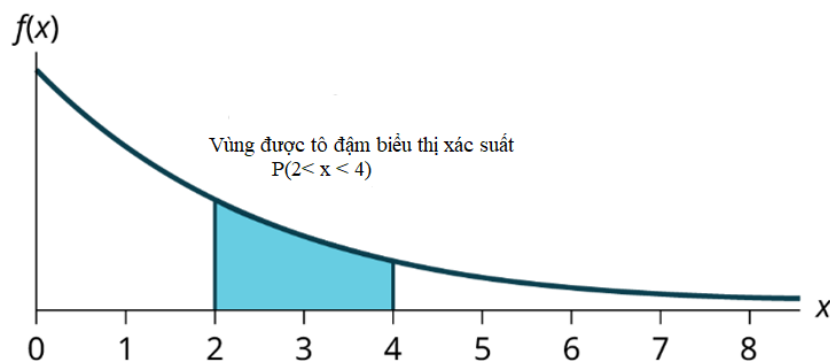
Có rất nhiều phân phối xác suất liên tục khác nhau. Khi dùng phân phối liên tục để mô hình hóa xác suất, chúng ta sẽ chọn phân phối phù hợp nhất với tình huống thực tế.

Trong chương này và chương sau, chúng ta sẽ tìm hiểu về: **phân phối đồng nhất** (Uniform Distribution), **phân phối mũ** (Exponential Distribution), **phân phối chuẩn** (Normal Distribution). Các biểu đồ sau đây sẽ minh họa cho những phân phối này.



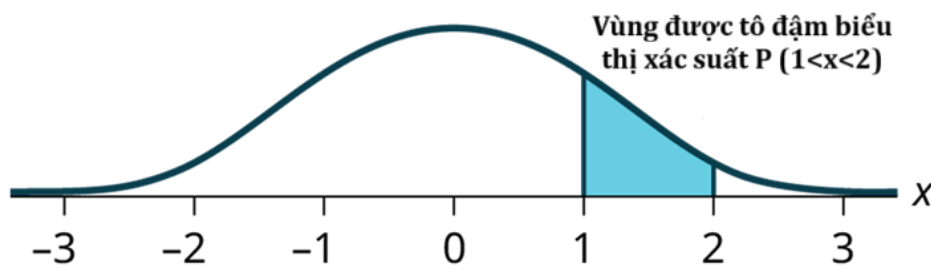
PHÂN PHỐI ĐỒNG NHẤT

Hình 5.2 Biểu đồ cho thấy một phân phối đồng nhất với vùng giữa $x = 3$ và $x = 6$ được tô đậm để biểu thị xác suất việc giá trị của biến ngẫu nhiên x nằm trong khoảng từ 3 đến 6.



PHÂN PHỐI MŨ

Hình 5.3 Biểu đồ cho thấy một phân phối mũ với vùng giữa $x = 2$ và $x = 4$ được tô đậm để biểu thị xác suất việc giá trị của biến ngẫu nhiên X nằm trong khoảng từ 2 đến 4.



PHÂN PHỐI CHUẨN

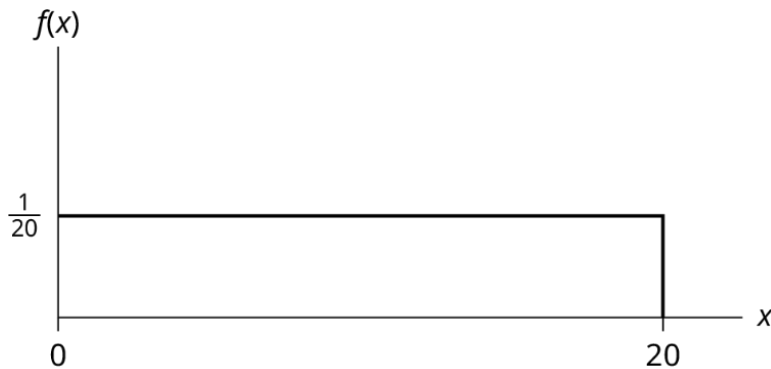
Hình 5.4 Biểu đồ cho thấy phân phối chuẩn với vùng giữa $x = 1$ và $x = 2$ được tô đậm để biểu thị xác suất việc giá trị của biến ngẫu nhiên x nằm trong khoảng từ 1 đến 2.

Đối với các phân phối xác suất liên tục, **XÁC SUẤT = DIỆN TÍCH**.

Ví dụ 5.1

Xét hàm số $f(x) = \frac{1}{20}$ trong khoảng $0 \leq x \leq 20$ với x là một số thực.

Đồ thị của $f(x) = \frac{1}{20}$ là một đường thẳng nằm ngang. Tuy nhiên, vì điều kiện $0 \leq x \leq 20$, nên hàm số $f(x)$ bị giới hạn trong đoạn từ $x = 0$ đến $x = 20$, bao gồm cả hai đầu mút.



Hình 5.5

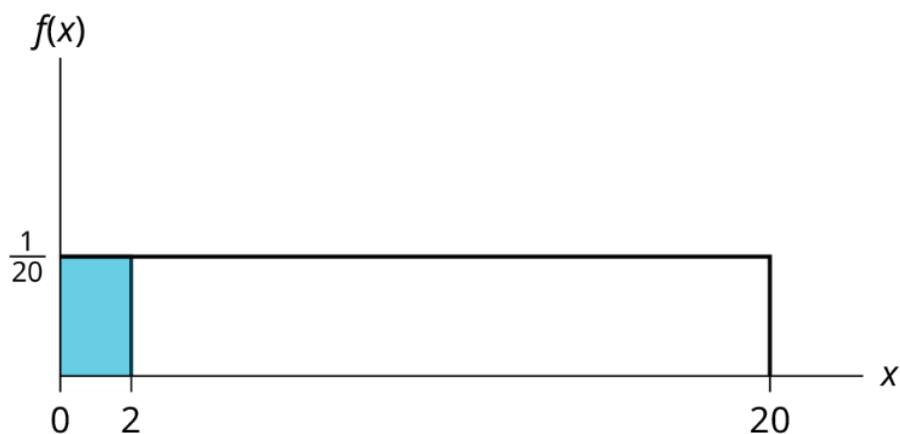
$f(x) = \frac{1}{20}$ với $0 \leq x \leq 20$

Đồ thị của $f(x) = \frac{1}{20}$ là một đoạn thẳng nằm ngang khi $0 \leq x \leq 20$.

Diện tích giữa $f(x) = \frac{1}{20}$ khi $0 \leq x \leq 20$ và trục x là diện tích của một hình chữ nhật có đáy $= 20$ và chiều cao $= \frac{1}{20}$.

$$\text{DIỆN TÍCH} = 20 \left(\frac{1}{20} \right) = 1$$

Giả sử chúng ta muốn tìm diện tích giữa $f(x) = \frac{1}{20}$ và trục x khi $0 < x < 2$



Hình 5.6

$$\text{DIỆN TÍCH} = (2 - 0) \left(\frac{1}{20} \right) = 0,1$$

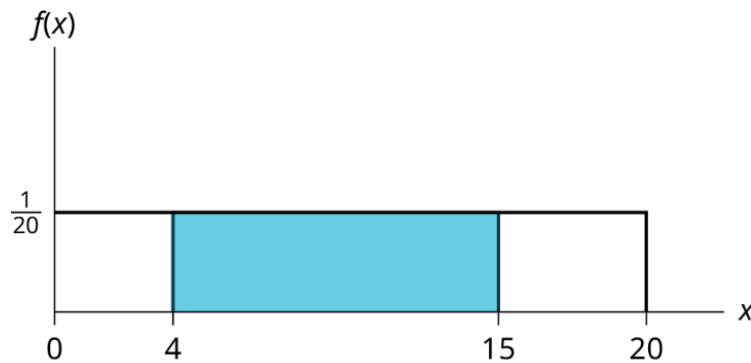
$(2 - 0) = 2 =$ chiều dài (đáy) của một hình chữ nhật

Nhắc lại

Diện tích hình chữ nhật = (chiều rộng) (chiều dài)

Phần diện tích trên tương ứng với một xác suất. Xác suất x nằm giữa 0 và 2 là 0,1, có thể được viết về mặt toán học là $P(0 < x < 2) = P(x < 2) = 0,1$.

Giả sử chúng ta muốn tìm diện tích giữa $f(x) = \left(\frac{1}{20}\right)$ và trục x tại đó $4 < x < 15$



Hình 5.7

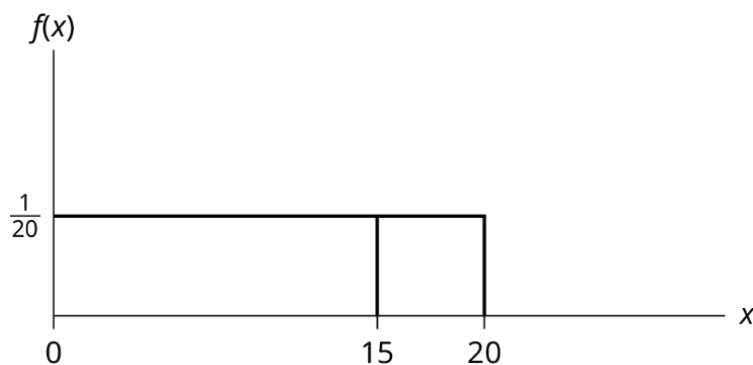
$$\text{DIỆN TÍCH} = (15 - 4) \left(\frac{1}{20}\right) = 0,55$$

$$(15 - 4) = 11 = \text{diện tích của một hình chữ nhật}$$

Diện tích tương ứng với xác suất $P(4 < x < 15) = 0,55$.

Giả sử chúng ta muốn tìm $P(x = 15)$. Trên đồ thị xy, $x = 15$ là một đường thẳng đứng. Đường thẳng đứng không có chiều rộng (hoặc chiều rộng bằng không).

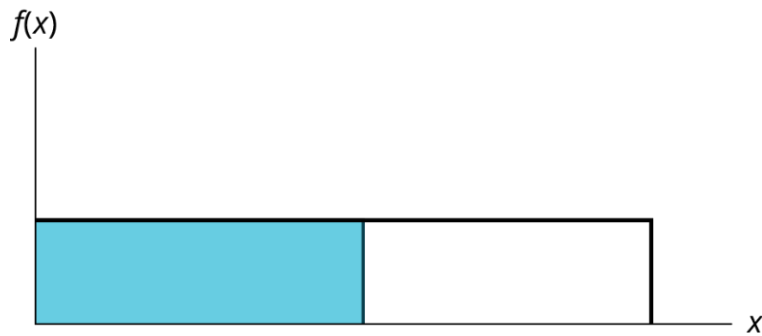
$$\text{Do đó, } P(x = 15) = (\text{dài đáy}) (\text{chiều cao}) = (0) \left(\frac{1}{20}\right) = 0$$



Hình 5.8

$P(X \leq x)$, cũng có thể được viết là $P(X < x)$ đối với phân phối liên tục, được gọi là hàm phân phối tích lũy hoặc CDF. Lưu ý ký hiệu "nhỏ hơn hoặc bằng". Chúng ta cũng có thể sử dụng CDF để tính $P(X > x)$. CDF đưa ra "diện tích bên trái" và $P(X > x)$ đưa ra "diện tích bên phải".

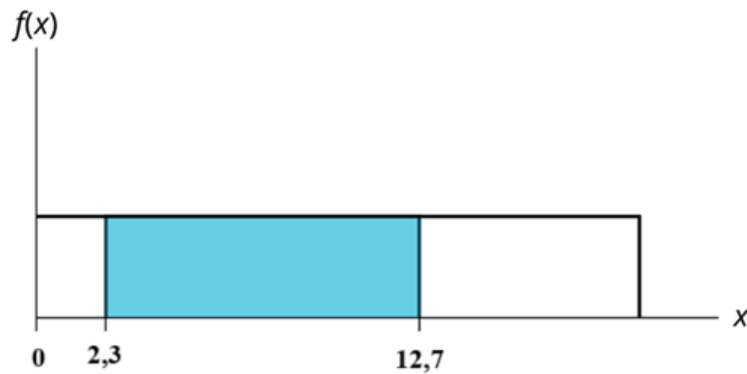
Chúng ta tính $P(X > x)$ đối với phân phối liên tục như sau: $P(X > x) = 1 - P(X < x)$.



Hình 5.9

Đánh dấu đồ thị bằng $f(x)$ và x . Chia tỷ lệ trục x và y theo giá trị lớn nhất của x và y , $f(x) = \left(\frac{1}{20}\right)$, $0 \leq x \leq 20$.

Để tính xác suất x nằm giữa hai giá trị, hãy xem đồ thị sau. Vùng được tô đậm giữa $x = 2,3$ và $x = 12,7$. Sau đó tính diện tích hình chữ nhật được tô đậm.



Hình 5.10

$$P(2,3 < x < 12,7) = (\text{dài đáy}) (\text{chiều cao}) = (12,7 - 2,3) \left(\frac{1}{20}\right) = 0,52$$

Bài tập vận dụng 5.1

Xét hàm số $f(x) = \frac{1}{8}$ với $0 \leq x \leq 8$. Vẽ đồ thị của $f(x)$ và tìm $P(2,5 < x < 7,5)$

5.2 Phân phối đồng nhất

Phân phối đồng nhất (uniform distribution) là phân phối xác suất liên tục, được dùng cho biến ngẫu nhiên liên quan tới các sự kiện có khả năng xảy ra như nhau. Khi giải quyết các bài toán có phân phối đồng nhất, hãy cẩn thận lưu ý xem dữ liệu có bao gồm hay không bao gồm các điểm cuối.

Phát biểu toán học của phân phối đồng nhất là:

$$f(x) = \frac{1}{b-a} \text{ với } a \leq x \leq b$$

trong đó a = giá trị thấp nhất của x và b = giá trị cao nhất của x .

Công thức cho số bình quân và độ lệch chuẩn lý thuyết là:

$$\mu = \frac{a+b}{2} \text{ và } \sigma = \sqrt{\frac{(b-a)^2}{12}}$$

Ví dụ 5.2

Khoảng thời gian (tính bằng phút) mà một người phải đợi xe buýt có phân phối đồng nhất giữa 0 và 15 phút.

Vấn đề

- Xác suất một người phải đợi ít hơn 12,5 phút là bao nhiêu?
- Trung bình, một người phải đợi trong bao lâu? Tìm số bình quân μ và độ lệch chuẩn σ .
- Trong 90% trường hợp, thời gian chờ đợi của một người sẽ ít hơn bao nhiêu? LƯU Ý: Câu này yêu cầu tìm bách phân vị thứ 90 (90th percentile).

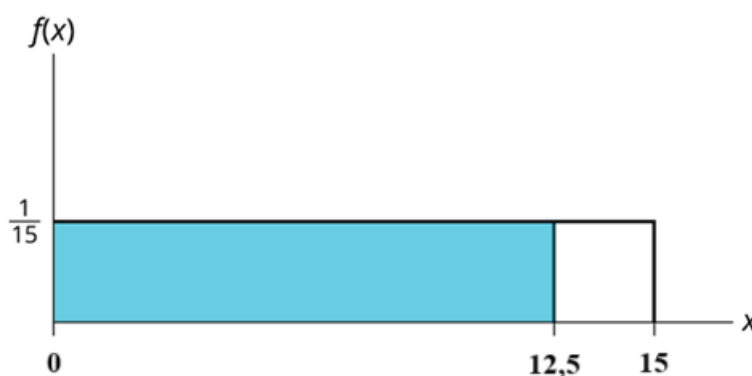
Lời giải

- Giả sử X = số phút mà một người phải đợi xe buýt, $a = 0$ và $b = 15$. $X \sim U(0, 15)$. Viết hàm mật độ xác suất $f(x) = \frac{1}{15-0} = \frac{1}{15}$ với $0 \leq x \leq 15$.

Tìm $P(x < 12,5)$. Vẽ đồ thị.

$$P(x < k) = (\text{dài đáy}) (\text{chiều cao}) = (12,5 - 0) \left(\frac{1}{15}\right) = 0,8333.$$

Xác suất một người phải đợi ít hơn 12,5 phút là 0,8333.



Hình 5.11

- $\mu = \frac{a+b}{2} = \frac{15+0}{2} = 7,5$. Trung bình, một người phải đợi 7,5 phút.

$$\sigma = \sqrt{\frac{(b-a)^2}{12}} = \sqrt{\frac{(15-0)^2}{12}} = 4,3. \text{ Độ lệch chuẩn là 4,3 phút.}$$

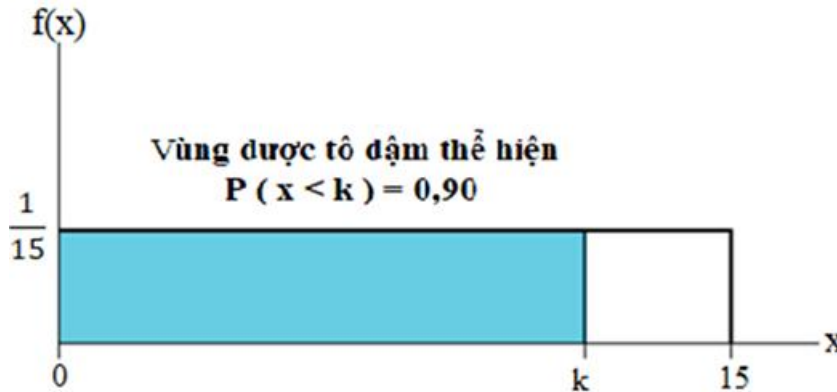
- Tìm bách phân vị thứ 90. Vẽ đồ thị. Giả sử k = bách phân vị thứ 90.

$$P(x < k) = (\text{dài đáy}) (\text{chiều cao}) = (k - 0) \left(\frac{1}{15}\right)$$

$$0,90 = (k) \left(\frac{1}{15} \right)$$

$$k = (0,90) (15) = 13,5$$

Bách phân vị thứ 90 là 13,5 phút. Chín mươi phần trăm trường hợp, một người phải đợi nhiều nhất là 13,5 phút.



Hình 5.12

LƯU Ý

Sử dụng Excel, R để phát số ngẫu nhiên theo phân phối đồng nhất.

Excel: đôi khi ta cần phát số ngẫu nhiên theo phân phối đồng nhất trong khoảng (a, b) . Sử dụng công thức $=\text{RAND}() * (b - a) + a$ sẽ cho một số ngẫu nhiên theo phân phối đồng nhất trong khoảng (a, b) .

R: hàm `runif(1, min = a, max = b)` sẽ tạo ra một số ngẫu nhiên theo phân phối đồng nhất trong khoảng (a, b) .

5.3 Phân phối mũ

Hàm mật độ xác suất **phân phối mũ** (exponential distribution) được xây dựng dựa trên **hàm số mũ tổng quát** (general exponential function), trong đó biến số nằm ở phần số mũ $f(x) = ab^x$. Phương trình này có thể được chuyển đổi sang hệ logarit tự nhiên với cơ số e (e xấp xỉ bằng 2,71828). Phân phối mũ có nhiều ứng dụng thực tiễn quan trọng, đặc biệt thường được dùng để mô tả khoảng thời gian chờ cho đến khi một sự kiện cụ thể xảy ra. Ví dụ, khoảng thời gian (tính từ hiện tại) cho đến khi xảy ra một trận động đất tuân theo phân phối mũ. Các ví dụ khác bao gồm thời lượng (tính bằng phút) của các cuộc gọi điện thoại đường dài hoặc thời gian sử dụng (tính bằng tháng) của một bình ắc-quy xe hơi. Người ta cũng chỉ ra rằng lượng tiền bạn mang theo trong ví hoặc túi xách thường xấp xỉ tuân theo phân phối mũ.

Các giá trị của biến ngẫu nhiên mũ thường xuất hiện theo cách sau: giá trị lớn ít khi xảy ra, giá trị nhỏ thường hay xảy ra. Ví dụ, các nghiên cứu tiếp thị cho thấy số tiền mà khách hàng chi trong một lần đi siêu thị tuân theo một phân phối mũ – phần lớn người tiêu dùng chi số tiền nhỏ và chỉ một số ít người chi số tiền lớn.

Phân phối mũ thường được sử dụng trong các tính toán liên quan đến độ tin cậy của sản phẩm, tức là khoảng thời gian sản phẩm có thể hoạt động. Một công ty sẽ dùng phân tích phân phối mũ để phân tích xác định số tháng bảo hành sản phẩm, dựa trên xác suất sản phẩm hỏng sau một khoảng thời gian cụ thể.

Biến ngẫu nhiên trong phân phối mũ là liên tục và thường đo lường sự trôi qua của thời gian, mặc dù cũng có thể áp dụng trong các tình huống biến ngẫu nhiên liên tục khác. Các câu hỏi thường gặp là: “Xác suất sự kiện xảy ra trong vòng x giờ hoặc ngày tới là bao nhiêu?” hoặc “Xác suất sự kiện xảy ra trong khoảng từ x_1 giờ đến x_2 giờ là bao nhiêu?” hoặc “Xác suất sự kiện cần hơn x_1 giờ để thực hiện là bao nhiêu?” Nói ngắn gọn, biến ngẫu nhiên X bằng (a) thời gian giữa hai sự kiện hoặc (b) khoảng thời gian để hoàn thành một hành động (ví dụ: phục vụ một khách hàng). Cần phân biệt biến ngẫu nhiên của phân phối mũ với phân phối Poisson. Phân phối Poisson là phân phối rời rạc, cho biết xác suất xảy ra x lần sự kiện trong một khoảng thời gian nhất định. Ngược lại, phân phối mũ là phân phối liên tục, cung cấp cho ta xác suất từ hiện tại và lần xảy ra tiếp theo của sự kiện sẽ có một khoảng thời gian dài X nào đó. Chúng ta sẽ khám phá thêm mối liên hệ giữa phân phối mũ và phân phối Poisson.

Hàm mật độ xác suất của phân phối mũ được cho bởi:

$$f(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}$$

Trong đó μ là thời gian chờ trung bình trong quá khứ.

Bình quân và độ lệch chuẩn của phân phối mũ đều bằng μ .

Để tìm xác suất của một biến ngẫu nhiên liên tục, ta tích phân hàm phân phối trong khoảng giá trị x tương ứng. Tích phân xác định của một hàm sẽ cho biết diện tích dưới đường cong hàm số nằm giữa hai giá trị của x .

Với phân phối mũ, có ba trường hợp tùy vào dạng câu hỏi.

Trường hợp I

Trường hợp I trả lời câu hỏi “Xác suất thời gian để hoàn tất nhỏ hơn x là bao nhiêu?” Câu hỏi xác suất sẽ được viết là: $P(X \leq x)$.

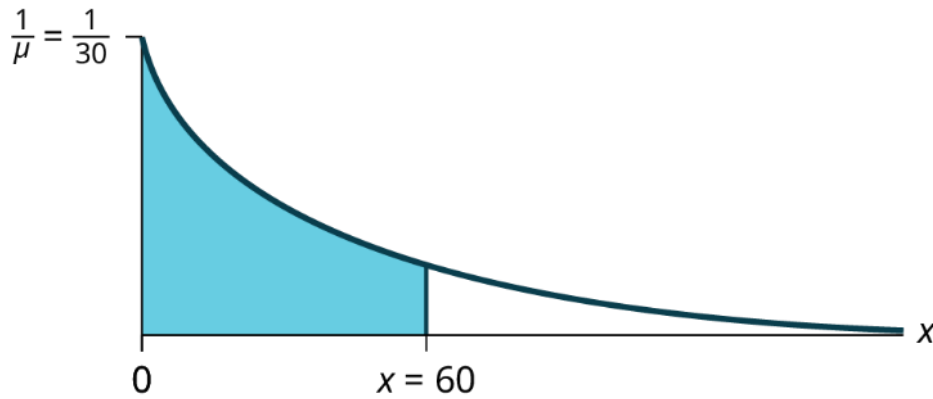
Ví dụ: Để tính xác suất một khách hàng sẽ được phục vụ trong vòng một giờ tới, biết rằng thời gian chờ trung bình là 30 phút, ta tích phân hàm mật độ xác suất phân phối mũ $\frac{1}{\mu} e^{-\frac{x}{\mu}}$ từ $x = 0$ đến $x = 60$.

$$F(x) = P(X \leq x) = \int_0^x \frac{1}{\mu} e^{-\frac{x}{\mu}} dx = 1 - e^{-\frac{x}{\mu}}$$

Phần diện tích bên dưới hàm mật độ xác suất mũ nằm giữa từ 0 đến 60 phút là xác suất cần tìm. Lưu ý rằng giá trị tại điểm 0 của hàm mật độ phân phối mũ là $\frac{1}{\mu}$. Giải phương trình hàm mật độ xác suất ở trên trong đó $\mu = 30$ là thời gian chờ trong quá khứ và $x = 60$ là thời gian chúng ta quan tâm đến.

$$P(X \leq 60) = \left(1 - e^{-\frac{x}{\mu}}\right) = 1 - e^{-\frac{60}{30}} = 0,865$$

Với thời gian chờ trong quá khứ là 30 phút, hiển nhiên ta tính được xác suất khách hàng sẽ phục vụ trong vòng một giờ tới là khoảng 0,865 – gần như chắc chắn, và được minh họa trong hình 5.13.



Hình 5.13

Trường hợp II

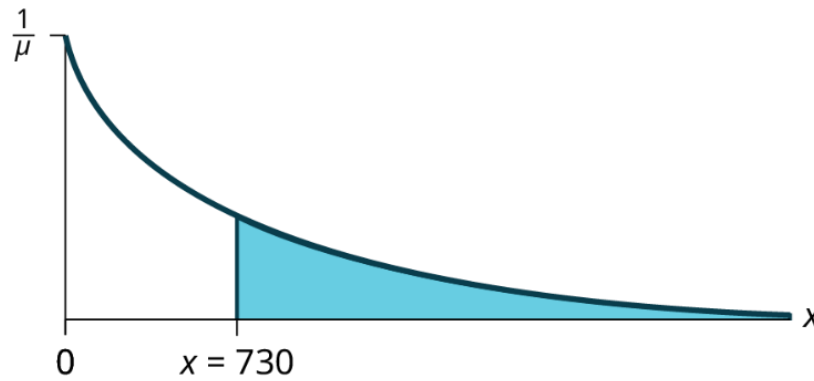
Trường hợp II trả lời câu hỏi: “Xác suất lớn hơn x là bao nhiêu?” Xác suất một người sẽ phải chờ NHIỀU HƠN x phút là bao nhiêu. Hàm xác suất liên quan sẽ là:

$$P(X \geq x) = 1 - \left(1 - e^{-\frac{x}{\mu}}\right) = e^{-\frac{x}{\mu}}$$

Giả sử một linh kiện điện tử được bảo hành trong vòng hai năm. Nếu linh kiện bị hỏng trong thời gian bảo hành hai năm này, người mua sẽ được hoàn tiền 100%. Nhà sản xuất, thông qua việc theo dõi cẩn thận các sản phẩm của mình, biết rằng tuổi thọ trung bình trong quá khứ của linh kiện này là 750 ngày. Liệu họ có đang thực hiện một chính sách kinh doanh rủi ro?

Ở đây, chúng ta có thể tính xác suất rằng một linh kiện sẽ hoạt động lâu hơn 730 ngày (hai năm), nghĩa là công ty không phải hoàn tiền. Lịch sử bình quân sử dụng là 750 ngày, và khoảng thời gian liên quan là $x = 730$ ngày.

Sử dụng hàm mật độ xác suất của phân phối mũ, ta thấy rằng xác suất để linh kiện điện tử này hoạt động vượt quá hai năm chỉ là 0,378, như minh họa trong Hình 5.14.



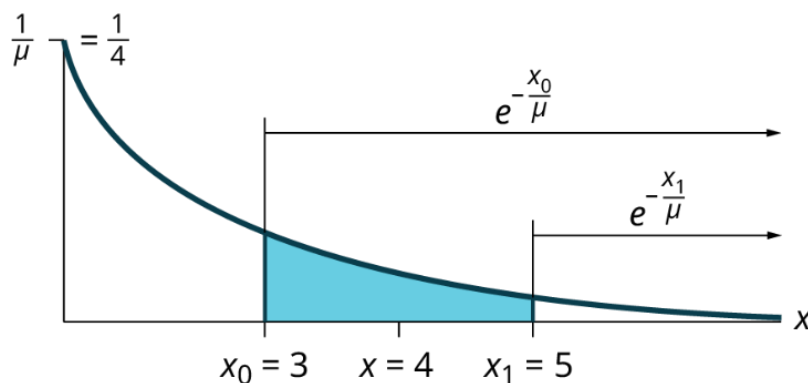
Hình 5.14

$$P(X \geq x) = e^{-\frac{x}{\mu}} = e^{-\frac{730}{750}} = 0,378$$

Vì tuổi thọ hoạt động của mỗi linh kiện điện tử là một biến ngẫu nhiên độc lập, nên ta có thể nói rằng khoảng 38% các linh kiện này có thể sử dụng hơn hai năm, đồng nghĩa với việc 62% người mua sẽ yêu cầu hoàn tiền toàn bộ. Nói ngắn gọn, chất lượng sản xuất thấp dẫn đến tỷ lệ hoàn tiền cao sẽ gây ra tổn thất tài chính đáng kể. Có thể dự đoán rằng chính sách bảo hành hai năm sẽ sớm bị loại bỏ.

Trường hợp III

Trường hợp III cho phép tính xác suất rằng một sự kiện sẽ xảy ra trong một khoảng thời gian cụ thể. Ví dụ, khi hỏi về thời gian đến của một chuyến bay, người ta có thể nhận được một khoảng thời gian như “cộng hoặc trừ so với thời gian đến dự kiến trong quá khứ được tính theo giờ kể từ thời điểm cất cánh”. Trong trường hợp này, số bình quân là thời gian đến dự kiến dựa trên dữ liệu lịch sử về thời gian di chuyển. Giả sử thời gian di chuyển trong quá khứ giữa hai điểm này là bốn giờ. Trường hợp III yêu cầu tính xác suất trong khoảng giữa hai điểm. Hình 5.15 minh họa Trường hợp III với hai điểm cách thời gian đến dự kiến một giờ sớm hơn và một giờ muộn hơn so với thời gian trung bình nằm ở giữa hai điểm đó.



Hình 5.15

Nếu thời gian di chuyển trong quá khứ là bốn giờ, thì x_0 thời gian di chuyển ba giờ và x_1 là thời gian di chuyển năm giờ. Biểu thức toán học là:

$$P(x_0 \leq x \leq x_1) = P(X \geq x_0) - P(X \geq x_1) = e^{-\frac{x_0}{\mu}} - e^{-\frac{x_1}{\mu}}$$

$$P(3 \leq x \leq 5) = e^{-\frac{3}{4}} - e^{-\frac{5}{4}} = 0,186$$

Xác suất để chuyến bay đến trong khoảng cộng hoặc trừ một giờ so với thời gian đến dự kiến là 4 giờ kể từ khi cất cánh là 0,186. Về mặt trực giác, đây là một xác suất rất nhỏ để đạt được một khoảng thời gian rộng như vậy. Cốt lõi của vấn đề là độ lệch chuẩn của thời gian bay trung bình phải rất lớn.

Phân phối mũ và tham số phân rã

Cốt lõi của phân phối mũ là e , cơ sở của logarit tự nhiên trong một hàm số mũ với biến số $-\frac{x}{\mu}$ nằm ở số mũ. Số mũ là số âm, do đó nó mô tả một hàm giảm theo cấp số mũ, như chúng ta đã thấy trong Hình 5.13, Hình 5.14 và Hình 5.15 ở trên.

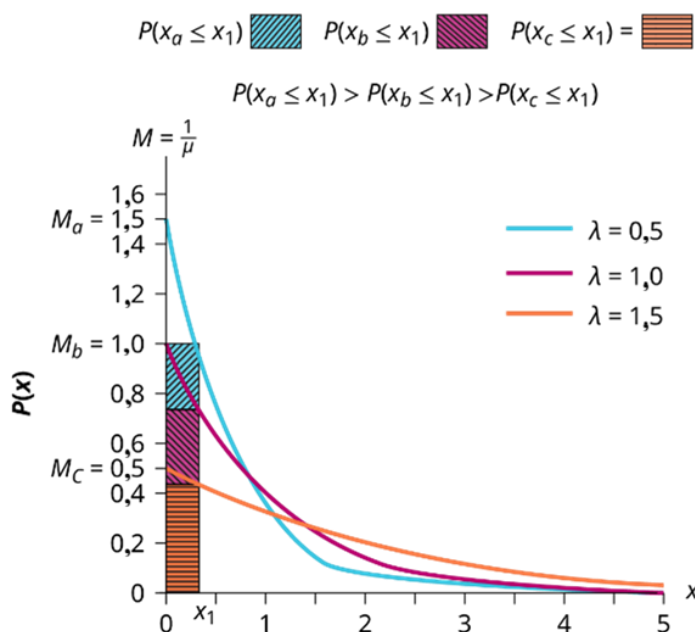
Điều này dẫn đến một công thức thay thế cho phân phối xác suất mũ:

$$f(x) = me^{-mx}$$

$$F(x) = \int_0^x me^{-mt} dt = 1 - e^{-mx}$$

Trong đó $m = \frac{1}{\mu}$ đo lường tốc độ suy giảm và được gọi là tham số phân rã (decay factor)¹².

Tham số phân rã đơn giản là đo mức độ nhanh chóng mà xác suất xảy ra của một sự kiện giảm xuống khi biến ngẫu nhiên X tăng lên.



Hình 5.16

¹² Còn gọi là hằng số phân rã, tham số tốc độ, và ký hiệu là λ .

Hình 5.16 minh họa phân phối mũ với ba tham số phân rã khác nhau. Như trước đây, điểm giao trên trục tung (mật độ xác suất) chính là tham số phân rã. Tham số phân rã bằng 1,5, trong hàm mũ đầu tiên, có tốc độ gấp ba lần phân phối mũ thứ ba với tham số phân rã là 0,5. Quan sát xác suất của mỗi hàm trong ba hàm này tại cùng một giá trị của x_1 , có thể thấy rõ rằng hàm đầu tiên có xác suất xảy ra cao hơn đáng kể so với hai hàm còn lại (tức là, $P(X_a \leq x_1) > P(X_b \leq x_1) > P(X_c \leq x_1)$, so sánh xác suất tương ứng của các hàm a, b, c theo thứ tự).

Cuối cùng, phân phối mũ cho thấy một cách rõ ràng ý nghĩa thực sự của “hàm số mũ”. Ở mỗi đơn vị, một sự thay đổi của x trong phân phối mũ dẫn đến một thay đổi đáng kể về xác suất. Ví dụ, nếu hàm phân phối mũ đơn giản $Y = 2^x$ khi thay đổi từ $x = 4$ đến $x = 5$, giá trị y tăng từ 16 lên 32, nhưng một thay đổi một đơn vị từ $x = 10$ sang $x = 11$ lại làm cho y tăng từ 1,024 lên 2,048. Tương tự, trong một hàm mũ với số mũ âm như phân phối xác suất mũ, sự thay đổi là rất đáng kể đối với mỗi đơn vị thay đổi tại các mức x cao hơn. Từ $x = -4$ đến $x = -5$, giá trị y giảm từ 0,063 xuống 0,031, trong khi một thay đổi một đơn vị từ $x = -10$ đến $x = -11$ làm giá trị y giảm từ 0,00097 xuống 0,00048.

Kết luận là: các xác suất trong một phân phối xác suất mũ tại các giá trị x lớn là rất hiếm, gần như không tồn tại.

LƯU Ý

Sử dụng Excel, R để tính phân phối mũ.

Excel: xác suất theo phân phối mũ $P(X \leq x)$ có thể tính bằng hàm EXPON.DIST($x, m, 1$); m là tham số phân rã.

R/stats package: xác suất theo phân phối mũ $P(X \leq x)$ có thể tính bằng hàm pexp(x, m); m là tham số phân rã.

Ví dụ 5.3

Giả sử x là khoảng thời gian (tính bằng phút) mà một nhân viên bưu điện dành cho một khách hàng. Theo dữ liệu lịch sử, thời gian trung bình dành cho mỗi khách hàng là 4 phút.

Ta biết rằng $\mu = 4$ phút, nghĩa là thời gian trung bình mà nhân viên dành cho một khách hàng là 4 phút. Hãy nhớ rằng chúng ta vẫn đang làm việc với xác suất, do đó cần được cung cấp các tham số quần thể như số bình quân. Để thực hiện bất kỳ phép tính nào, ta cần biết số bình quân của phân phối: ví dụ như thời gian phục vụ theo dữ liệu lịch sử. Việc biết số bình quân lịch sử cho phép tính tham số phân rã, m .

$$m = \frac{1}{\mu}. \text{ Vì vậy, } m = \frac{1}{4} = 0,25$$

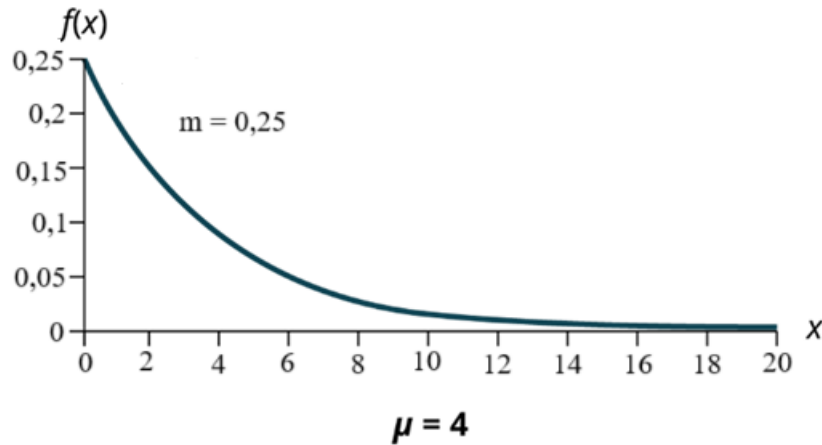
Khi ký hiệu tham số phân rã là m , hàm mật độ xác suất được trình bày như sau $f(x) = me^{-mx}$, đó đơn giản là công thức gốc với m thay thế cho $\frac{1}{\mu}$, hay $f(x) = \frac{1}{\mu} e^{-\frac{1}{\mu}x}$.

Để tính xác suất phân phối mũ, chúng ta cần sử dụng hàm mật độ xác suất. Như được thể hiện dưới đây, đường cong của hàm mật độ xác suất là:

$f(x) = 0,25e^{-0,25x}$ trong đó $x \geq 0$ và $m = 0,25$.

Ví dụ $f(5) = 0,25e^{(-0,25)(5)} = 0,072$. Nói cách khác, hàm có giá trị là 0,072 khi $x = 5$.

Biểu đồ như sau:



Hình 5.17

Lưu ý rằng đồ thị là một đường cong giảm dần. Khi $x = 0$:

$$f(x) = 0,25e^{(-0,25)(0)} = (0,25)(1) = 0,25 = m.$$

Giá trị lớn nhất trên trục tung (trục y) luôn là m , tức là một chia cho số bình quân.

Bài tập vận dụng 5.3

Thời gian mà các cặp vợ chồng dành để chọn thiệp kỷ niệm có thể được mô hình hóa bằng phân phối mũ với thời gian trung bình là tám phút. Hãy viết biểu thức của phân phối, nêu hàm mật độ xác suất, và vẽ đồ thị của phân phối này.

Ví dụ 5.4

Vấn đề

Sử dụng thông tin trong Ví dụ 5.3, hãy tìm xác suất rằng một nhân viên dành từ bốn đến năm phút với một khách hàng được chọn ngẫu nhiên.

Lời giải

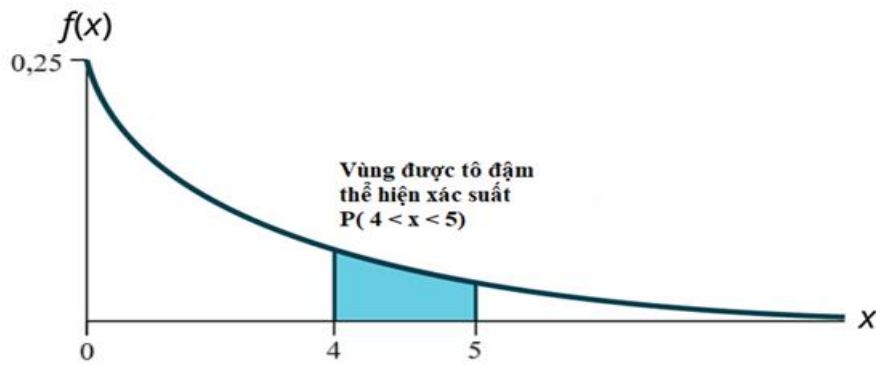
Tìm $P(4 < x < 5)$

Hàm phân phối tích lũy (CDF) cho biết diện tích bên trái của đồ thị.

$$P(x < X) = 1 - e^{-mx}.$$

$$P(x < 5) = 1 - e^{(-0,25)(5)} = 0,7135 \text{ và } P(x < 4) = 1 - e^{(-0,25)(4)} = 0,6321$$

$$P(4 < x < 5) = 0,7135 - 0,6321 = 0,0814$$



Hình 5.8

Bài tập vận dụng 5.4

Số ngày trước khi hành khách mua vé máy bay có thể được mô hình hóa bằng phân phối mũ với thời gian trung bình là 15 ngày. Hãy tìm xác suất rằng một hành khách sẽ mua vé trước **ít hơn mười ngày**. Bao nhiêu ngày là thời gian mà **một nửa** số hành khách chờ?

Ví dụ 5.5

Trung bình, một bộ phận máy tính nhất định có tuổi thọ là mười năm. Thời gian mà bộ phận máy tính này hoạt động tuân theo phân phối mũ.

Vấn đề

- Xác suất, một bộ phận máy tính nhất định có tuổi thọ là mười năm.
- Trung bình, năm bộ phận máy tính sẽ hoạt động trong bao lâu nếu chúng được sử dụng lần lượt, cái này sau cái kia?
- Xác suất rằng một bộ phận máy tính hoạt động trong khoảng từ 9 đến 11 năm là bao nhiêu?

Lời giải

- Gọi X là thời gian (tính bằng năm) mà bộ phận máy tính hoạt động.

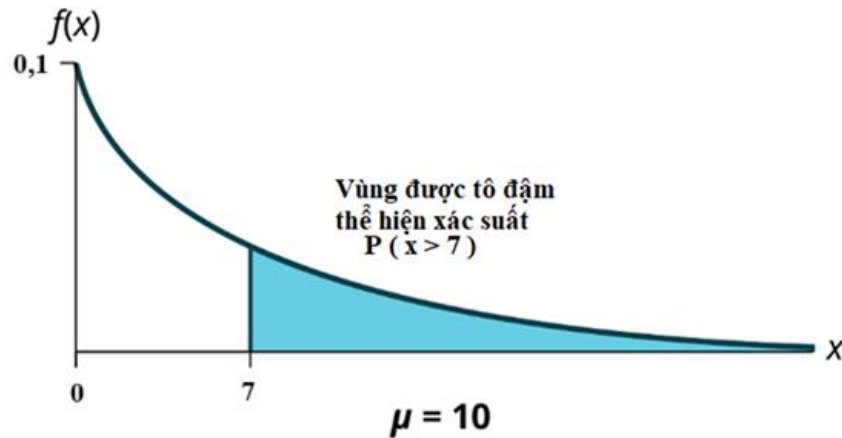
Ta có $\mu = 10$ do đó $m = \frac{1}{\mu} = \frac{1}{10} = 0,1$.

Tìm $P(X > 7)$. Vẽ đồ thị.

$$P(X > 7) = 1 - P(X < 7).$$

$$\text{Vì } P(X < x) = 1 - e^{-mx} \text{ nên } P(X > x) = 1 - (1 - e^{-mx}) = e^{-mx}.$$

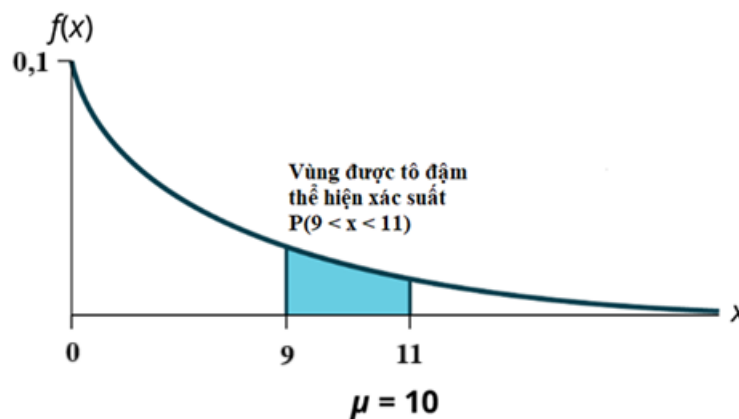
$P(X > 7) = e^{(-0,1)(7)} = 0,4966$. Vậy xác suất rằng một bộ phận máy tính hoạt động được hơn 7 năm là 0,4966.



Hình 5.19

b. Trung bình, một bộ phận máy tính hoạt động trong 10 năm. Do đó, nếu năm bộ phận được sử dụng lần lượt, thì tổng thời gian trung bình sẽ là: $(5)(10) = 50$ năm.

c. Tìm $P(9 < X < 11)$. Vẽ đồ thị.



Hình 5.20

$P(9 < X < 11) = P(X < 11) - P(X < 9) = (1 - e^{(-0,1)(11)}) - (1 - e^{(-0,1)(9)}) = 0,6671 - 0,5934 = 0,0737$. Xác suất một bộ phận máy tính hoạt động trong khoảng từ 9 đến 11 năm là 0,0737.

Bài tập vận dụng 5.5

Trung bình, một đôi giày chạy bộ có thể sử dụng được trong 18 tháng nếu được sử dụng hằng ngày. Thời gian sử dụng của giày chạy bộ tuân theo phân phối mũ. Xác suất rằng một đôi giày chạy bộ có thể sử dụng được hơn 15 tháng là bao nhiêu? Trung bình, sáu đôi giày chạy bộ sẽ sử dụng được bao lâu nếu dùng lần lượt, đôi này sau đôi kia? Tám mươi phần trăm số giày chạy bộ sẽ sử dụng được tối đa bao lâu nếu dùng hằng ngày?

Ví dụ 5.6

Giả sử rằng thời lượng của một cuộc gọi điện thoại (tính bằng phút) là một biến ngẫu nhiên phân phối mũ với tham số phân rã $\frac{1}{12}$. Nếu có người khác đến trạm điện thoại công cộng ngay trước bạn, hãy tìm xác suất rằng bạn sẽ phải chờ **hơn năm phút**. Gọi X là thời lượng của một cuộc gọi điện thoại, tính bằng phút.

Vấn đề

Giá trị của m , μ , và σ là bao nhiêu? Xác suất rằng bạn phải chờ hơn năm phút là _____.

Lời giải

$$m = \frac{1}{12}$$

$$\mu = 12$$

$$\sigma = 12$$

$$P(X > 5) = 0,6592$$

Bài tập vận dụng 5.6

Giả sử rằng khoảng cách (tính bằng dặm) mà mọi người sẵn sàng đi để đến nơi làm việc là một biến ngẫu nhiên phân phối mũ với một **tham số phân rã $\frac{1}{20}$** . Gọi X là khoảng cách (tính bằng dặm) mà mọi người sẵn sàng đi để đi làm. Giá trị của m , μ , và σ là bao nhiêu? Xác suất rằng một người sẵn sàng đi làm với khoảng cách **hơn 25 dặm** là bao nhiêu?

Ví dụ 5.7

Thời gian chờ giữa các sự kiện thường được mô hình hóa bằng phân phối mũ. Ví dụ, giả sử rằng trung bình có 30 khách hàng đến cửa hàng mỗi giờ và thời gian giữa các lần đến tuân theo phân phối mũ.

Vấn đề

- Trung bình có bao nhiêu phút trôi qua giữa hai lượt khách đến liên tiếp?
- Khi cửa hàng vừa mở cửa, trung bình mất bao lâu để có ba khách hàng đến?
- Sau khi một khách hàng đến, tìm xác suất rằng mất **dưới một phút** để khách hàng tiếp theo đến.
- Sau khi một khách hàng đến, tìm xác suất rằng mất **hơn năm phút** để khách hàng tiếp theo đến.
- Liệu phân phối mũ có hợp lý cho tình huống này không?

Lời giải

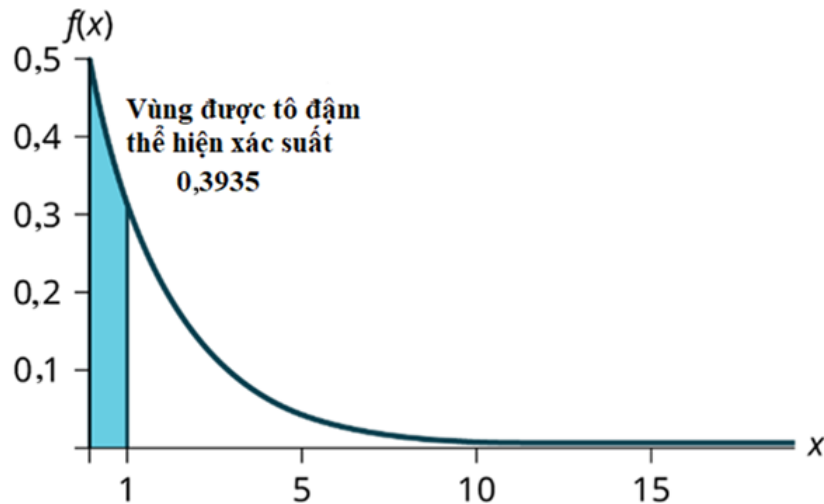
- Vì chúng ta kỳ vọng 30 khách hàng đến mỗi giờ (tức 60 phút), nên kỳ vọng trung bình sẽ có một khách hàng đến **mỗi 2 phút**.

b. Vì trung bình một khách hàng đến mỗi 2 phút, nên sẽ mất trung bình **6 phút** để có 3 khách hàng đến.

c. Gọi X là thời gian giữa các lượt khách đến, tính bằng phút. Theo phần a, $\mu = 2$ nên $m = \frac{1}{2} = 0,5$.

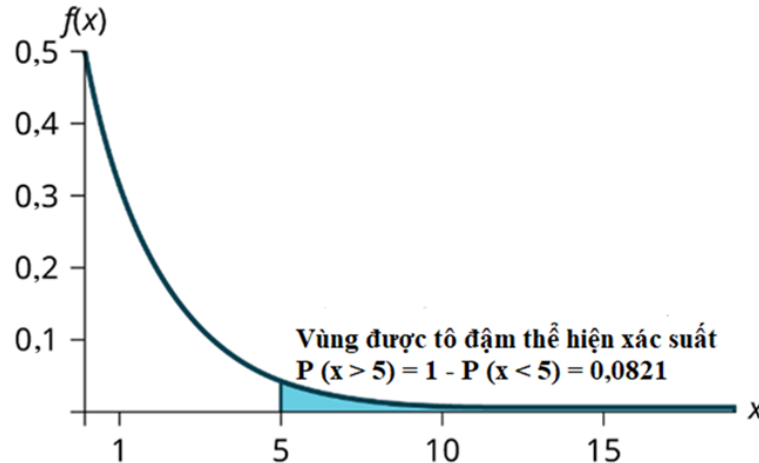
Hàm phân phối tích lũy (CDF) là: $P(X < x) = 1 - e^{(-0,5)(x)}$

Vì vậy $P(X < 1) = 1 - e^{(-0,5)(1)} = 0,3935$.



Hình 5.21

d. $P(X > 5) = 1 - P(X < 5) = 1 - e^{(-0,5)(5)} = e^{-2,5} \approx 0,0821$.



Hình 5.22

e. Mô hình này giả định rằng mỗi lần chỉ có một khách hàng đến, điều này có thể không hợp lý vì mọi người có thể đi mua sắm theo nhóm, dẫn đến nhiều khách hàng đến cùng lúc. Mô hình cũng giả định rằng dòng khách hàng không thay đổi trong suốt cả ngày, điều này không đúng nếu có những thời điểm trong ngày đông khách hơn những thời điểm khác.

Bài tập vận dụng 5.7

Giả sử rằng trên một đoạn đường cao tốc nhất định, xe ô tô đi qua với mật độ trung bình là năm xe mỗi phút. Giả định rằng khoảng thời gian giữa các xe liên tiếp tuân theo phân phối mũ.

- Trung bình có bao nhiêu giây trôi qua giữa hai xe liên tiếp?
- Sau khi một chiếc xe đi qua, trung bình mất bao lâu để bảy chiếc xe tiếp theo đi qua?
- Tìm xác suất rằng sau khi một chiếc xe đi qua, chiếc xe tiếp theo sẽ đi qua **trong vòng 20 giây tới**.
- Tìm xác suất rằng sau khi một chiếc xe đi qua, sẽ **không có chiếc xe nào đi qua trong ít nhất 15 giây** tiếp theo.

Tính chất không nhớ của phân phối mũ

Trong ví dụ 5.7 khoảng thời gian giữa hai lần khách hàng đến cửa hàng được mô hình hóa theo phân phối mũ với một số bình quân là hai phút. Giả sử đã trôi qua năm phút kể từ khi khách hàng cuối cùng đến. Vì một khoảng thời gian khá dài đã trôi qua, ta có thể nghĩ rằng khả năng có khách hàng đến trong phút tiếp theo sẽ cao hơn. Tuy nhiên, với phân phối mũ, điều này không đúng — khoảng thời gian chờ thêm cho đến khách hàng tiếp theo không phụ thuộc vào thời gian đã trôi qua kể từ lần khách trước. Đây chính là Tính chất không nhớ (memoryless property). Phân phối mũ và phân phối hình học là hai loại phân phối duy nhất có Tính chất không nhớ này. Cụ thể, **Tính chất không nhớ** (memoryless property) phát biểu rằng:

$$P(X > r + t \mid X > r) = P(X > t) \text{ với mọi } r \geq 0 \text{ và } t \geq 0$$

Ví dụ, nếu đã trôi qua 5 phút kể từ khi khách hàng cuối cùng đến, thì xác suất rằng sẽ còn hơn một phút nữa cho đến khi khách tiếp theo đến được tính bằng cách sử dụng $r = 5$ và $t = 1$ trong phương trình trên:

$$P(X > 5 + 1 \mid X > 5) = P(X > 1) = e^{(-0,5)(1)} = 0,6065.$$

Đây chính là xác suất cho việc phải chờ hơn một phút để có một khách hàng mới đến tiếp theo sau bất kỳ khách nào trước đó.

Phân phối mũ thường được dùng để mô hình hóa tuổi thọ của một thiết bị điện hoặc cơ khí. Trong Ví dụ 5.5, tuổi thọ của một bộ phận máy tính tuân theo phân phối mũ với trung bình là 10 năm. **Tính chất không nhớ** chỉ ra rằng những gì đã xảy ra trong quá khứ không ảnh hưởng đến xác suất trong tương lai. Trong trường hợp này, điều đó có nghĩa là một bộ phận cũ không có khả năng hỏng hóc cao hơn tại bất kỳ thời điểm nào so với một bộ phận mới. Nói cách khác, bộ phận hoạt động như mới cho đến khi nó đột ngột hỏng. Ví dụ, nếu một bộ phận đã hoạt động được 10 năm, thì xác suất nó hoạt động thêm được 7 năm nữa là:

$$P(X > 17 \mid X > 10) = 0,4966, \text{ trong đó ký hiệu "}" được đọc là "cho trước", hoặc "nếu như".}$$

Ví dụ 5.8

Quay lại ví dụ về nhân viên bưu điện, trong đó thời gian mà một nhân viên bưu điện dành cho một khách hàng tuân theo phân phối mũ với trung bình là bốn phút. Giả sử một khách hàng đã được phục vụ bốn phút. Vậy xác suất rằng khách hàng đó sẽ tiếp tục được phục vụ ít nhất thêm ba phút nữa là bao nhiêu?

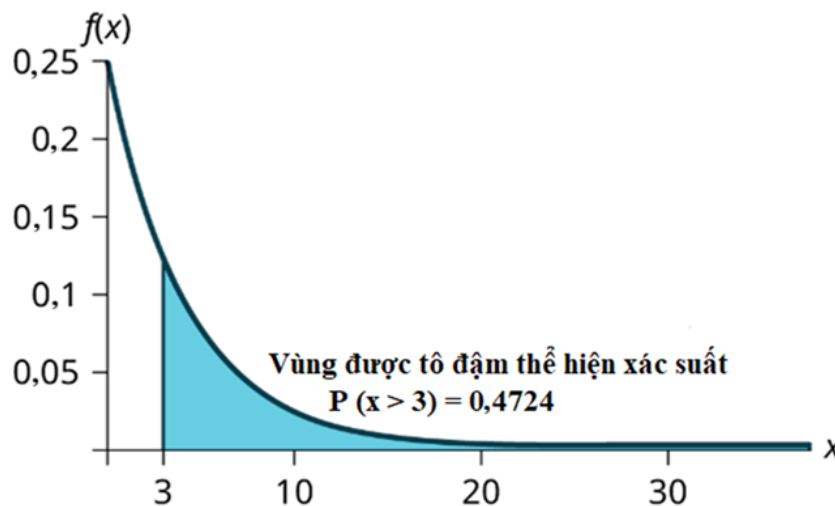
Tham số phân rã của X là $m = \frac{1}{4} = 0,25$, do đó $X \sim \text{Exp}(0,25)$.

Hàm phân phối tích lũy (CDF) là: $P(X < x) = 1 - e^{-0,25x}$.

Ta cần tìm $P(X > 7 \mid X > 4)$.

Tính chất không nhớ cho ta biết rằng: $P(X > 7 \mid X > 4) = P(X > 3)$. Vì vậy, ta chỉ cần tìm xác suất rằng một khách hàng sẽ được phục vụ hơn ba phút nữa bởi nhân viên bưu điện.

Đó là $P(X > 3) = 1 - P(X < 3) = 1 - (1 - e^{-(0,25)(3)}) = e^{-0,75} \approx 0,4724$.



Hình 5.23

Bài tập vận dụng 5.8

Giả sử rằng tuổi thọ của một bóng đèn tuân theo phân phối mũ với tuổi thọ trung bình là **tám năm**. Nếu một bóng đèn đã sử dụng được **12 năm**, hãy tìm xác suất rằng nó sẽ sử dụng được **tổng cộng hơn 19 năm**.

Mối quan hệ giữa phân phối Poisson và phân phối mũ

Có một mối quan hệ thú vị giữa phân phối mũ và phân phối Poisson. Giả sử rằng khoảng thời gian trôi qua giữa hai sự kiện liên tiếp tuân theo phân phối mũ với một số bình quân theo thời gian là μ . Cũng giả định rằng các khoảng thời gian này độc lập với nhau, tức là thời gian giữa các sự kiện không bị ảnh hưởng bởi những khoảng thời gian giữa các sự kiện trước đó.

Nếu các giả định này đúng, thì số lượng sự kiện xảy ra trong một đơn vị thời gian sẽ tuân theo phân phối Poisson với số bình quân là μ . Ta biết rằng nếu X tuân theo phân phối Poisson với số bình quân μ , thì $P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$.

Trong khi đó công thức của phân phối mũ: $P(X = x) = me^{-mx} = \frac{1}{\mu} e^{-\frac{1}{\mu}x}$, với m là tham số phân rã, μ là thời gian trung bình giữa các lần xảy ra.

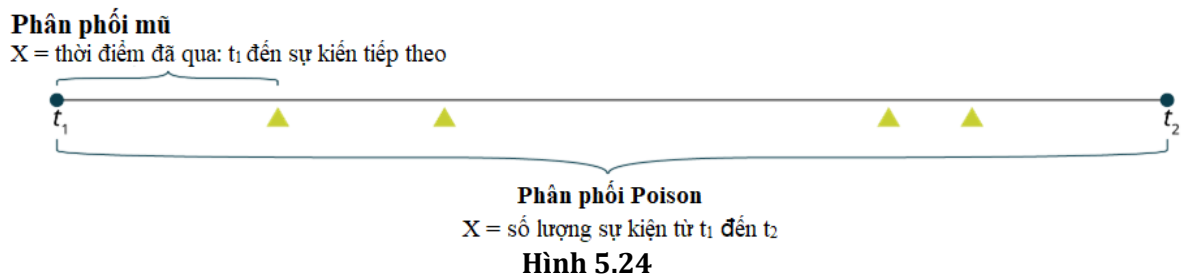
Ta thấy rằng phân phối mũ là "họ hàng" với phân phối Poisson, và chúng liên kết với nhau thông qua công thức trên. Tuy nhiên, có những khác biệt quan trọng khiến mỗi phân phối phù hợp với các loại bài toán xác suất khác nhau.

Trước hết, phân phối Poisson là phân phối với biến ngẫu nhiên rời rạc X , trong đó thời gian là một biến liên tục là được chia nhỏ thành các phần rời rạc. Như ta đã biết, **số lần xảy ra** của một sự kiện trong một khoảng thời gian xác định, x , tuân theo phân phối Poisson. Ví dụ: Số lần điện thoại đổ chuông trong một giờ.

Ngược lại, **khoảng thời gian** giữa các lần xảy ra sự kiện tuân theo phân phối mũ. Ví dụ: Điện thoại vừa đổ chuông, mất bao lâu nữa để nó đổ chuông lại? Ở đây ta đo độ dài thời gian giữa các sự kiện, là một biến ngẫu nhiên liên tục sử dụng phân phối mũ, không phải số lần xảy ra sự kiện trong khoảng sử dụng phân phối Poisson.

Phân phối mũ so với phân phối Poisson

Một cách trực quan để thể hiện cả sự tương đồng và khác biệt giữa hai phân phối này là dùng một dòng thời gian.



Biến ngẫu nhiên trong phân phối Poisson là rời rạc, đếm số lần xảy ra sự kiện trong một khoảng thời gian xác định, từ t_1 đến t_2 như minh họa trong Hình 5.24, và hàm phân phối Poisson tính xác suất xảy ra số lần xuất hiện đó. Số lần xảy ra sự kiện (4 lần trong đồ thị) được đo bằng số đếm vì vậy biến ngẫu nhiên trong phân phối Poisson là **biến rời rạc**.

Phân phối xác suất mũ thì tính xác suất dựa trên **thời gian trôi qua** - một **biến ngẫu nhiên liên tục**. Trong Hình 5.24, điều này được thể hiện bằng đoạn dấu ngoặc từ t_1 đến thời điểm sự kiện tiếp theo xảy ra được đánh dấu bằng hình tam giác.

Các câu hỏi kinh điển cho phân phối **Poisson** là: “Sẽ có bao nhiêu người đến quầy thanh toán của tôi trong giờ tới?”

Các câu hỏi kinh điển cho phân phối **mũ** là: “Sẽ mất bao lâu nữa cho đến khi người tiếp theo đến?” hoặc một biến thể: “Người đó sẽ ở lại quầy bao lâu sau khi đến?”

Một lần nữa, công thức của phân phối mũ là:

$$f(x) = me^{-mx} \text{ hoặc } f(x) = \frac{1}{\mu} e^{-\frac{1}{\mu}x}$$

Ta có thể thấy ngay sự tương đồng giữa công thức phân phối mũ và phân phối Poisson.

$$P(x) = \frac{\mu^x e^{-\mu}}{x!}$$

Cả hai hàm phân phối xác suất đều dựa trên mối quan hệ giữa thời gian và sự tăng trưởng hoặc suy giảm theo cấp số mũ. Chữ “e” trong công thức là một hằng số có giá trị xấp xỉ 2,71828 và là cơ sở của hàm logarit tự nhiên — công thức nền tảng cho tăng trưởng theo hàm mũ. Khi người ta nói thứ gì đó “tăng trưởng theo cấp số mũ”, thì đó chính là điều đang được nói đến.

Ví dụ về phân phối mũ và Poisson sẽ làm rõ sự khác biệt giữa hai phân phối này và đồng thời cho thấy ứng dụng thú vị của chúng.

Phân phối Poisson

Giả sử rằng **trong quá khứ**, có 10 khách hàng đến quầy thanh toán mỗi giờ. Hãy nhớ rằng đây vẫn là bài toán xác suất, nên **các giá trị quá khứ phải được cung cấp**. Ta nhận thấy đây là một **bài toán phân phối Poisson**.

Ta có thể đưa thông tin này vào hàm mật độ xác suất Poisson để có một **công thức tổng quát** dùng để tính xác suất cho bất kỳ số lượng khách cụ thể nào sẽ đến trong giờ tiếp theo.

Công thức có thể áp dụng được với **bất kỳ giá trị nào của biến ngẫu nhiên rời rạc được chọn, và khi đó** x được đặt vào công thức. Trong ví dụ này, $\mu = 10$ vì ta kỳ vọng 10 khách mỗi giờ, nên đưa $\mu = 10$ vào công thức.

$$f(x) = \frac{10^x e^{-10}}{x!}$$

Tiếp theo, ví dụ **xác suất có 15 người đến quầy thanh toán trong giờ tới** là:

$$P(x = 15) = \frac{10^{15} e^{-10}}{15!} = 0,0347$$

Sau khi thay số $x = 15$ và tính toán, ta được kết quả **xác suất có 15 người đến trong giờ tới là 0,035**.

Phân phối mũ

Nếu ta **giữ nguyên dữ liệu quá khứ** rằng có 10 khách hàng đến mỗi giờ, nhưng bây giờ **quan tâm đến thời gian phục vụ của một người tại quầy**, thì ta sử dụng **phân phối mũ**. Hàm xác suất mũ cho bất kỳ giá trị nào của x (biến ngẫu nhiên liên tục) với dữ liệu quá khứ này là:

$$f(x) = \frac{1}{0,1} e^{-x/0,1} = 10e^{-10x}$$

Để tính μ , **thời gian phục vụ trung bình**, ta đơn giản chia số người đến trong mỗi giờ, 10, trong khoảng thời gian một giờ và có $\mu = 0,1$ nghĩa là **trung bình mỗi người mất 0,1 giờ hoặc 6 phút** tại quầy. Đây chính là lý do tại sao có số 0,1 trong công thức.

Có một sự **nhầm lẫn phổ biến** với ký hiệu μ trong cả hai công thức của Poisson và mũ. Mặc dù ký hiệu giống nhau, nhưng **ý nghĩa của μ là khác nhau**:

- Trong **Poisson**, μ là **số sự kiện trung bình xảy ra** trong một đơn vị thời gian.
- Trong **phân phối mũ**, μ là **thời gian trung bình giữa các sự kiện**.

Chúng có **mối quan hệ nghịch đảo**.

Tiếp tục với ví dụ tại quầy thanh toán; nếu ta muốn biết **xác suất rằng một người sẽ mất không quá 9 phút để thanh toán**, thì:

- Đổi đơn vị: 9 phút = 0,15 giờ
- Vì đang hỏi về một khoảng giá trị (dưới 9 phút), và đây là **biến liên tục**, nên câu hỏi xác suất là: $P(x \leq 9) = 1 - 10e^{-10x}$.

Thay số vào công thức, ta tính được: $P(x = 0,15) = 1 - 10e^{-10(0,15)} = 0,7769$.

Vậy xác suất một khách hàng sẽ rời khỏi quầy trong vòng 9 phút là 0,7769.

Ta thấy rằng **xác suất rời đi trong vòng 9 phút là khá cao**, trong khi **xác suất có 15 khách hàng đến trong giờ tiếp theo là rất nhỏ (0,035)**.

Ví dụ 5.9

Tại một đồn cảnh sát ở một thành phố lớn, số cuộc gọi đến với tốc độ trung bình là **bốn cuộc gọi mỗi phút**. Lưu ý rằng chúng ta **chỉ quan tâm đến mật độ các cuộc gọi đến**, và **bỏ qua thời gian thực hiện cuộc gọi**. Ta cũng cần giả định rằng **khoảng thời gian giữa các cuộc gọi là độc lập**. Điều này có nghĩa là một khoảng thời gian chờ dài giữa hai cuộc gọi **không làm tăng khả năng** cuộc gọi tiếp theo đến sớm hơn. Từ đó, ta có thể suy luận rằng **tổng số cuộc gọi nhận được trong một khoảng thời gian nhất định** sẽ tuân theo **phân phối Poisson**.

Vấn đề

- Tìm **thời gian trung bình giữa hai cuộc gọi liên tiếp**.
- Tìm **xác suất** rằng sau khi một cuộc gọi được nhận, cuộc gọi tiếp theo sẽ đến **trong vòng chưa đầy 10 giây**.
- Tìm **xác suất có đúng 5 cuộc gọi xảy ra trong một phút**.
- Tìm **xác suất có ít hơn 5 cuộc gọi xảy ra trong một phút**.
- Tìm **xác suất có nhiều hơn 40 cuộc gọi xảy ra trong vòng 8 phút**.

Lời giải

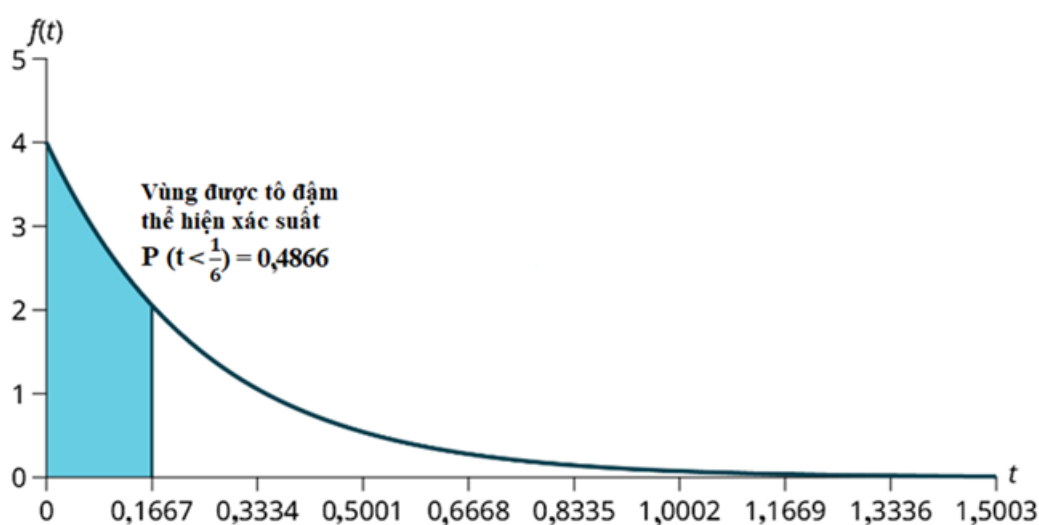
a. Trung bình có 4 cuộc gọi xảy ra mỗi phút $\rightarrow$ tức là một cuộc gọi xảy ra **mỗi 15 giây** hay là 0,25 phút là **khoảng thời gian trung bình giữa hai cuộc gọi liên tiếp**.

b. Gọi T là **thời gian giữa hai cuộc gọi**. Từ phần a, ta có $\mu = 0,25$ vì vậy $m = 4$. Do đó $T \sim \text{Exp}(4)$.

Hàm phân phối tích lũy là $P(T < t) = 1 - e^{-4(t)}$.

Xác suất cuộc gọi tiếp theo xảy ra trong vòng chưa đầy mười giây (mười giây = $\frac{1}{6}$ phút) là

$$P\left(T < \frac{1}{6}\right) = 1 - e^{-4\left(\frac{1}{6}\right)} = 0,4866.$$



Hình 5.25

Cho X = số cuộc gọi mỗi phút. Như đã nêu trước đó, số cuộc gọi mỗi phút có phân phối Poisson, với giá trị bình quân là bốn cuộc gọi mỗi phút.

Do đó, $X \sim \text{Poisson}(4)$ và do đó

$$P(X = 5) = \frac{4^5 e^{-4}}{5!} \approx 0,1563 \quad 5! = (5)(4)(3)(2)(1)$$

d. Hãy nhớ rằng X phải là một số nguyên, do đó $P(X < 5) = P(X \leq 4)$.

Để tính toán điều này, ta có thể:

$$P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3) + P(X = 4)$$

Sử dụng phần mềm để tính toán, chúng ta thấy rằng $P(X \leq 4) = 0,6288$

e. Cho Y = số cuộc gọi xảy ra trong khoảng thời gian tám phút.

Vì có trung bình bốn cuộc gọi mỗi phút, nên có trung bình $(8)(4) = 32$ cuộc gọi trong mỗi khoảng thời gian tám phút.

Do đó, $Y \sim \text{Poisson}(32)$. Do đó, $P(Y > 40) = 1 - P(Y \leq 40) = 1 - 0,9294 = 0,0707$.

Bài tập vận dụng 5.9

Tại một thành phố nhỏ, số vụ tai nạn ô tô xảy ra tuân theo phân phối Poisson với trung bình là 3 vụ mỗi tuần.

- Tính xác suất rằng có nhiều nhất 2 vụ tai nạn xảy ra trong một tuần bất kỳ.
- Xác suất rằng có ít nhất hai tuần giữa hai vụ tai nạn bất kỳ là bao nhiêu?

Thuật ngữ chính

Phân phối đồng nhất (Uniform Distribution): một biến ngẫu nhiên liên tục (RV) có các kết quả xảy ra đồng đều trên miền $a < x < b$. Nó cũng thường được gọi là phân phối hình chữ nhật vì đồ thị hàm mật độ xác suất là hình chữ nhật. Nó có bình quân $\mu = \frac{a+b}{2}$ và độ lệch chuẩn $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. Hàm mật độ xác suất $f(x) = \frac{1}{b-a}$ với $a < x < b$ hoặc $a \leq x \leq b$. Hàm phân phối tích lũy: $P(X \leq x) = \frac{x-a}{b-a}$.

Phân phối mũ (Exponential Distribution): một biến ngẫu nhiên liên tục (RV) dùng để mô tả khoảng thời gian giữa các sự kiện ngẫu nhiên — ví dụ như khoảng thời gian giữa các ca cấp cứu đến bệnh viện. Số bình quân μ , và độ lệch chuẩn $\sigma = \mu = \frac{1}{m}$. Hàm mật độ xác suất: $f(x) = me^{-mx}$ hoặc $f(x) = \frac{1}{\mu} e^{-\frac{1}{\mu}x}$, $x \geq 0$, và hàm phân phối tích lũy: $P(X \leq x) = 1 - e^{-mx}$ hoặc $P(X \leq x) = 1 - e^{-\frac{1}{\mu}x}$.

Phân phối Poisson (Poisson distribution): nếu có trung bình μ sự kiện xảy ra trong một đơn vị thời gian, và các sự kiện này là độc lập, thì số sự kiện X xảy ra trong một đơn vị thời gian sẽ tuân theo phân phối Poisson. Xác suất để có x sự kiện trong một đơn vị thời gian được tính bởi: $P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$.

Tham số phân rã (decay parameter): mô tả tốc độ mà xác suất giảm về 0 khi giá trị x tăng. Đây là giá trị m trong hàm mật độ xác suất $f(x) = m e^{-mx}$ của một biến ngẫu nhiên phân phối mũ. $m = \frac{1}{\mu}$, trong đó μ là số bình quân của biến ngẫu nhiên đó.

Tính chất không nhớ (memoryless property): Với biến ngẫu nhiên phân phối mũ X , thuộc tính không ghi nhớ phát biểu rằng những gì đã xảy ra trong quá khứ không ảnh hưởng đến xác suất trong tương lai. Điều này có nghĩa là xác suất để biến X lớn hơn $(x + t)$, biết rằng X lớn hơn x , sẽ giống như xác suất biến X lớn hơn t trong điều kiện ta không thông tin nào khác. Bằng ký hiệu, chúng ta nói rằng $P(X > x + t | X > x) = P(X > t)$.

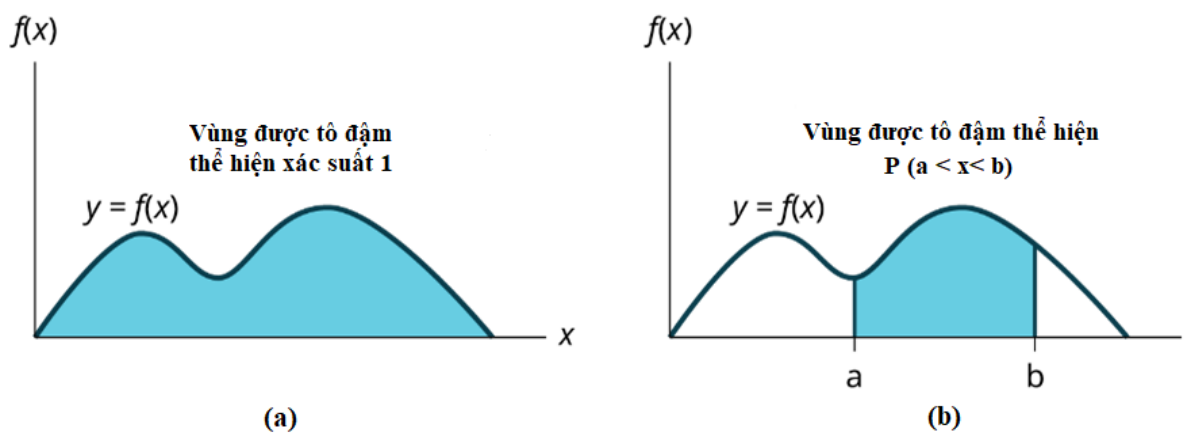
Xác suất có điều kiện (conditional probability): khả năng xảy ra của một sự kiện khi một sự kiện khác đã xảy ra.

Tóm tắt chương

5.1 Các tính chất của hàm mật độ xác suất liên tục

Hàm mật độ xác suất (probability density function - pdf) được sử dụng để mô tả xác suất của các biến ngẫu nhiên liên tục. Diện tích dưới đường cong mật độ nằm giữa hai điểm tương ứng với xác suất rằng biến ngẫu nhiên rơi vào khoảng giữa hai giá trị đó. Nói cách khác, diện tích dưới đường cong giữa điểm a và b bằng với xác suất $P(a < x < b)$.

Hàm phân phối tích lũy (CDF) thể hiện xác suất dưới dạng diện tích. Nếu X là một biến ngẫu nhiên liên tục, thì hàm mật độ xác suất $f(x)$ được sử dụng để vẽ đồ thị của phân phối xác suất. Tổng diện tích dưới đồ thị của $f(x)$ là bằng 1. Diện tích dưới đồ thị của $f(x)$ nằm giữa hai giá trị a và b chính là xác suất $P(a < x < b)$.



Hình 5.26

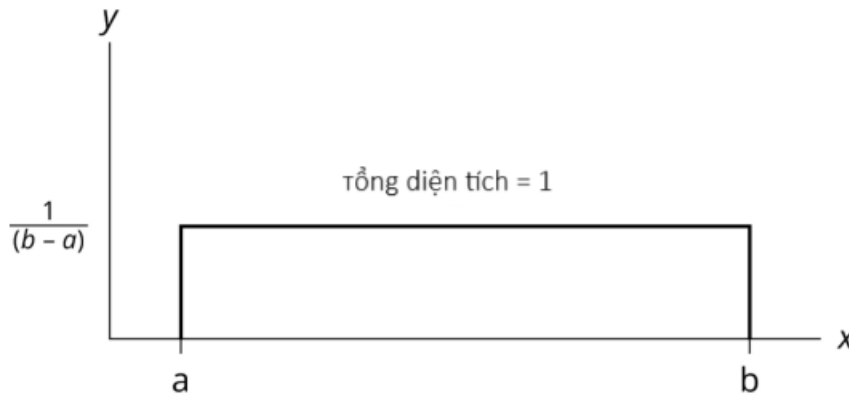
Hàm phân phối tích lũy (CDF) của biến ngẫu nhiên X được định nghĩa là $P(X \leq x)$. Đây là một hàm theo biến x , biểu diễn xác suất rằng biến ngẫu nhiên nhận giá trị **nhỏ hơn hoặc bằng x** .

5.2 Phân phối đồng nhất

Nếu X có phân phối đồng nhất trên khoảng $a < x < b$ hay $a \leq x \leq b$, thì X nhận các giá trị nằm trong khoảng từ a đến b (có thể bao gồm cả a và b). Tất cả các giá trị x trong khoảng đó đều có **xác suất như nhau**. Ta ký hiệu $X \sim U(a, b)$. Bình quân của X là $\mu = \frac{a+b}{2}$. Độ lệch

chuẩn của X là $\sigma = \sqrt{\frac{(b-a)^2}{12}}$. **Hàm mật độ xác suất** của X là $f(x) = \frac{1}{b-a}$ với $a \leq x \leq b$.

Hàm phân phối tích lũy của X là $P(X \leq x) = \frac{x-a}{b-a}$. Biến X là **liên tục**.



Hình 5.27

Xác suất $P(c < X < d)$ có thể được tính bằng cách lấy **diện tích dưới đường** $f(x)$ trong khoảng từ c đến d . Vì trong phân phối đồng nhất, diện tích tương ứng là một **hình chữ nhật**, nên ta chỉ cần **nhân chiều rộng với chiều cao** để tính diện tích.

5.3 Phân phối mũ

Nếu X là một biến ngẫu nhiên có **phân phối mũ** với số bình quân μ , thì **tham số** phân rã là $m = \frac{1}{\mu}$. **Hàm mật độ xác suất** của X là $f(x) = me^{-mx}$, (hoặc tương đương $f(x) = \frac{1}{\mu} e^{-\frac{1}{\mu}x}$). **Hàm phân phối tích lũy** của X là $P(X \leq x) = 1 - e^{-mx}$.

Tóm tắt công thức

5.1 Các tính chất của hàm mật độ xác suất liên tục

Hàm mật độ xác suất (PDF) $f(x)$

- $f(x) \geq 0$
- Tổng diện tích dưới đồ thị của $f(x)$ là bằng 1

5.2 Phân phối đồng nhất

X = một số thực nằm giữa a và b (trong một số trường hợp, X có thể nhận giá trị a và b). $a = X$ nhỏ nhất; $b = X$ lớn nhất.

$X \sim U(a, b)$.

Bình quân của X là $\mu = \frac{a+b}{2}$

Độ lệch chuẩn của X là $\sigma = \sqrt{\frac{(b-a)^2}{12}}$

Hàm mật độ xác suất của X là $f(x) = \frac{1}{b-a}$ với $a \leq x \leq b$

Diện tích bên trái của x : $P(X < x) = (x - a) \left(\frac{1}{b-a} \right)$

Diện tích bên phải của x : $P(X > x) = (b - x) \left(\frac{1}{b-a} \right)$

Diện tích giữa c và d : $P(c < x < d) = (\text{base})(\text{height}) = (d - c) \left(\frac{1}{b-a} \right)$

- pdf: $f(x) = \frac{1}{b-a}$ với $a \leq x \leq b$
- cdf: $P(X \leq x) = \frac{x-a}{b-a}$
- Bình quân $\mu = \frac{a+b}{2}$
- Độ lệch chuẩn $\sigma = \sqrt{\frac{(b-a)^2}{12}}$
- $P(c < X < d) = (d - c) \left(\frac{1}{b-a} \right)$

5.3 Phân phối mũ

- pdf: $f(x) = me^{-mx}$ trong đó $x \geq 0$ và $m > 0$
- cdf: $P(X \leq x) = 1 - e^{-mx}$
- Bình quân $\mu = \frac{1}{m}$
- Độ lệch chuẩn $\sigma = \frac{1}{m}$
- Ngoài ra
 - $P(X > x) = e^{-mx}$
 - $P(a < x < b) = e^{-ma} - e^{-mb}$
- Phân phối poisson: $P(X = x) = \frac{\mu^x e^{-\mu}}{x!}$ Với bình quân và phương sai của μ

Tài liệu tham khảo

5.2 Phân phối đồng nhất

McDougall, John A. The McDougall Program for Maximum Weight Loss. Plume, 1995.

5.3 Phân phối mũ

Data from the United States Census Bureau.

Data from World Earthquakes, 2013. Available online at <http://www.world-earthquakes.com/> (accessed June 11, 2013).

“No-hitter.” Baseball-Reference.com, 2013. Available online at <http://www.baseball-reference.com/bullpen/No-hitter> (accessed June 11, 2013).

Zhou, Rick. “Exponential Distribution lecture slides.” Available online at www.public.iastate.edu/~riczw/stat330s11/lecture/lec13.pdf (accessed June 11, 2013).

6 PHÂN PHỐI CHUẨN



Hình 6.1: Nếu bạn hỏi một số lượng đủ nhiều người về kích cỡ giày của họ, bạn sẽ thấy rằng dữ liệu có thể vẽ thành đồ thị có hình dạng giống như một đường cong hình chuông và có thể được mô tả là tuân theo phân phối chuẩn. (nguồn ảnh: chỉnh sửa từ tác phẩm “Shoe shop” của Lo/ Flickr, phạm vi công cộng)

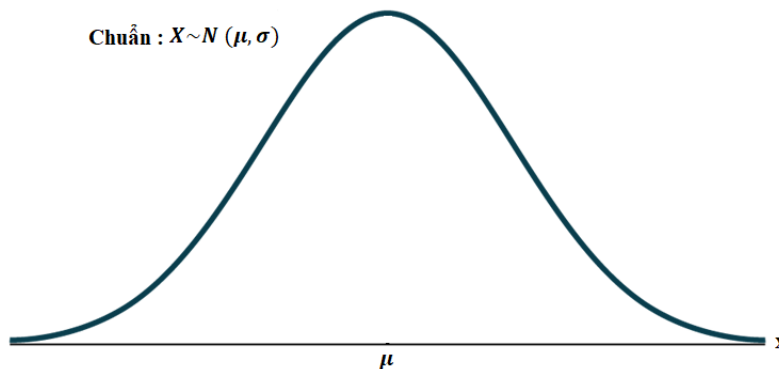
Giới thiệu

Hàm mật độ xác suất chuẩn (normal probability density function) - một phân phối liên tục, là hàm phân phối quan trọng nhất trong tất cả các loại phân phối. Nó được sử dụng rất phổ biến. Đồ thị của nó có dạng hình chuông. Bạn sẽ thấy đường cong hình chuông trong hầu hết mọi lĩnh vực. Một trong các lĩnh vực này là tâm lý học, kinh doanh, kinh tế học, khoa học tự nhiên, điều dưỡng và tất nhiên là toán học. Một số giảng viên của bạn có thể sử dụng phân phối chuẩn để giúp xác định điểm số của bạn. Hầu hết các điểm IQ là tuân theo phân phối chuẩn. Thậm chí giá bất động sản trong nhiều khu vực cũng thường có dạng phân phối chuẩn.

Phân phối chuẩn cực kỳ quan trọng, tuy nhiên nó không thể được áp dụng cho mọi thứ trong thế giới thực. Lưu ý rằng ở đây chúng ta vẫn đang nói về phân phối dữ liệu quần thể. Đây là một cuộc thảo luận về xác suất và do đó, nếu dữ liệu quần thể tuân theo phân phối chuẩn thì đây là cách chúng ta có thể tìm xác suất của các sự kiện cụ thể, giống như cách chúng ta đã làm đối với dữ liệu quần thể có thể có phân phối nhị thức hoặc phân phối Poisson. Lưu ý này được nêu lên ở đây vì trong chương tiếp theo chúng ta sẽ thấy rằng phân phối chuẩn mô tả một số điều rất khác từ dữ liệu thô và tạo thành nền tảng của thống kê suy luận.

Phân phối chuẩn có hai tham số (hai số đo mô tả): bình quân (μ) và độ lệch chuẩn (σ). Nếu X là một đại lượng được đo lường, có phân phối chuẩn với bình quân (μ) và độ lệch chuẩn

(σ), chúng ta biểu thị điều này bằng cách viết công thức của hàm mật độ xác suất chuẩn như sau:



Hình 6.2

Hàm mật độ xác suất chuẩn là một hàm toán học khá phức tạp. Bạn không cần phải ghi nhớ công thức này. Điều đó không cần thiết.

$$f(x) = \frac{1}{\sigma \cdot \sqrt{2 \cdot \pi}} e^{-\frac{1}{2} \left(\frac{x-\mu}{\sigma} \right)^2}$$

Đường cong của phân phối chuẩn có tính đối xứng quanh một đường thẳng đứng đi qua bình quân μ . Bình quân cũng chính là trung vị và bằng với yếu vị, vì đồ thị phân phối chuẩn có tính đối xứng quanh bình quân μ . Theo như ký hiệu đã thể hiện, phân phối chuẩn chỉ phụ thuộc vào bình quân và độ lệch chuẩn. Lưu ý rằng điều này khác với một số hàm mật độ xác suất mà bạn đã được học trước đó, chẳng hạn như Poisson, trong đó bình quân bằng μ và độ lệch chuẩn đơn giản là căn bậc hai của bình quân, hoặc phân phối nhị thức, trong đó p được sử dụng để xác định cả bình quân và độ lệch chuẩn. Vì diện tích dưới đường cong của hàm mật độ xác suất phải bằng một, nên sự thay đổi về độ lệch chuẩn σ sẽ gây ra sự thay đổi hình dạng của đường cong chuẩn; đường cong trở nên phình ra và rộng hơn, hoặc hẹp lại và cao hơn tùy thuộc vào σ . Sự thay đổi về μ khiến đồ thị dịch chuyển sang trái hoặc phải mà không làm thay đổi hình dạng của nó. Điều này có nghĩa là có vô số phân phối xác suất chuẩn khác nhau trong thực tế. Một phân phối đặc biệt được quan tâm được gọi là phân phối chuẩn đơn vị (standard normal distribution, unit normal distribution)¹³.

6.1. Phân phối chuẩn đơn vị

Phân phối chuẩn đơn vị (standard normal distribution) là phân phối chuẩn cho các giá trị chuẩn hóa có tên là **điểm chuẩn hóa z** (z-score). Một điểm chuẩn hóa z được đo lường bằng đơn vị độ lệch chuẩn.

Phân phối chuẩn đơn vị là có bình quân bằng 0 và độ lệch chuẩn bằng 1. Điều này làm đơn giản hóa đáng kể phép tính toán học về xác suất. Hãy dành một chút thời gian và thay thế

¹³ Một số tài liệu sử dụng từ “phân phối chuẩn tắc” để chỉ “standard normal distribution”, tuy nhiên “chuẩn tắc” cũng hay được dùng để dịch từ “normative” với ý nghĩa “quy chuẩn, quy tắc được khuyến cáo phải tuân theo” như trong từ “normative economics”.

0 và 1 vào những vị trí thích hợp trong công thức trên và bạn có thể thấy rằng phương trình thu gọn thành một phương trình có thể giải dễ dàng hơn nhiều bằng phép tính tích phân. Cho một giá trị x có phân phối chuẩn với bình quân μ và độ lệch chuẩn σ . Phép biến đổi $z = \frac{x - \mu}{\sigma}$ sẽ tạo ra các điểm chuẩn hóa z có phân phối $Z \sim N(0, 1)$. Điểm chuẩn hóa z cho biết khoảng cách từ một giá trị x cụ thể tới điểm bình quân μ bằng bao nhiêu độ lệch chuẩn σ .

Điểm chuẩn hóa z

Nếu X là một biến ngẫu nhiên tuân theo phân phối chuẩn $X \sim N(\mu, \sigma)$, thì điểm chuẩn hóa z (z -score) của một giá trị cụ thể x được tính theo công thức:

$$z = \frac{x - \mu}{\sigma}$$

Điểm chuẩn hóa z cho bạn biết cần bao nhiêu độ lệch chuẩn σ từ giá trị x nằm phía trên (về bên phải trục) hay phía dưới (về bên trái) đến số bình quân μ . Các giá trị x lớn hơn số bình quân có điểm z là số dương, và các giá trị x nhỏ hơn số bình quân có điểm z là số âm. Nếu x bằng số bình quân ($x = \mu$), thì x có điểm z bằng không.

Ví dụ 6.1

Giả sử $X \sim N(5, 6)$. Điều này cho thấy rằng X là một biến phân phối chuẩn ngẫu nhiên với bình quân $\mu = 5$ và độ lệch chuẩn $\sigma = 6$. Giả sử $x = 17$ thì:

$$z = \frac{x - \mu}{\sigma} = \frac{17 - 5}{6} = 2$$

Điều này có nghĩa là $x = 17$ **cách hai độ lệch chuẩn** (two standard deviations) (2σ) phía trên hoặc về bên phải của số bình quân $\mu = 5$.

Bây giờ, giả định $x = 1$. Thì $z = \frac{x - \mu}{\sigma} = \frac{1 - 5}{6} = -0,67$ (làm tròn đến hai chữ số thập phân)

Điều này có nghĩa rằng $x = 1$ cách 0,67 độ lệch chuẩn phía dưới ($-0,67\sigma$) hoặc về bên trái của số bình quân $\mu = 5$.

Bài tập vận dụng 6.1

Điền vào các chỗ trống.

Jerome ghi trung bình 16 điểm trong một trận đấu với độ lệch chuẩn là bốn điểm. $x \sim N(16, 4)$. Giả sử Jerome ghi được 10 điểm trong một trận đấu. Điểm chuẩn hóa z cho $x = 10$ là $-1,5$. Điểm này cho biết rằng $x = 10$ là ____ độ lệch chuẩn của ____ (bên phải hoặc bên trái) của số bình quân ____ (Số bình quân là bao nhiêu?).

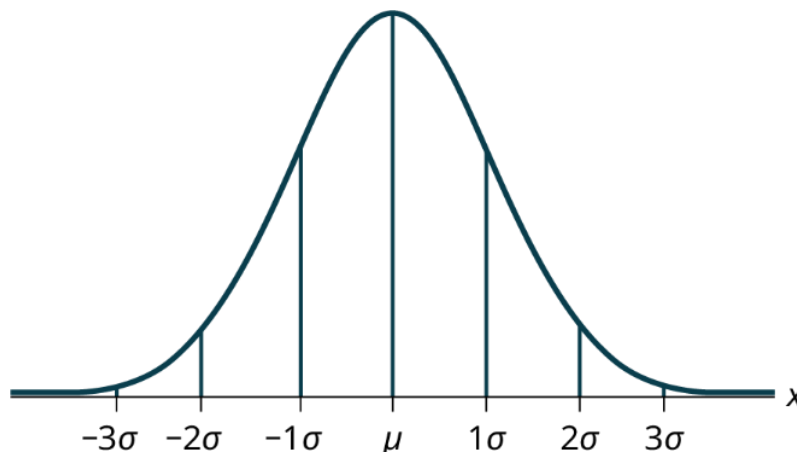
Quy tắc thực nghiệm

Nếu X là một biến ngẫu nhiên và có phân phối chuẩn với bình quân μ và độ lệch chuẩn σ , thì **Quy tắc thực nghiệm** (empirical rule) phát biểu như sau:

- Khoảng 68% giá trị x nằm trong khoảng từ -1σ đến $+1\sigma$ của bình quân μ (trong phạm vi một độ lệch chuẩn của bình quân).
- Khoảng 95% giá trị x nằm trong khoảng từ -2σ đến $+2\sigma$ của bình quân μ (trong phạm vi hai độ lệch chuẩn của bình quân).
- Khoảng 99,7% giá trị x nằm trong khoảng từ -3σ đến $+3\sigma$ của bình quân μ (trong phạm vi ba độ lệch chuẩn của bình quân).

LƯU Ý rằng hầu hết các giá trị x nằm trong phạm vi giữa ba độ lệch chuẩn quanh số bình quân.

- Điểm chuẩn hóa z cho $+1\sigma$ và -1σ là tương ứng $+1$ và -1 .
- Điểm chuẩn hóa z cho $+2\sigma$ và -2σ là tương ứng $+2$ và -2 .
- Điểm chuẩn hóa z cho $+3\sigma$ và -3σ là tương ứng $+3$ và -3 .



Hình 6.3

Ví dụ 6.2

Giả sử X có một phân phối chuẩn với bình quân 50 và độ lệch chuẩn 6.

Khoảng 68% của giá trị x nằm trong một độ lệch chuẩn của bình quân. Do đó, khoảng 68% giá trị x nằm giữa $-1\sigma = (-1)(6) = -6$ và $1\sigma = (1)(6) = 6$ của số bình quân 50. Các giá trị $50 - 6 = 44$ và $50 + 6 = 56$ nằm trong một độ lệch chuẩn so với số bình quân 50. Điểm z lần lượt là -1 và $+1$ tương ứng cho 44 và 56.

Khoảng 95% giá trị x nằm trong hai độ lệch chuẩn của bình quân. Do đó, khoảng 95% giá trị x nằm giữa $-2\sigma = (-2)(6) = -12$ và $2\sigma = (2)(6) = 12$. Các giá trị $50 - 12 = 38$ và $50 + 12 = 62$ nằm trong hai độ lệch chuẩn so với số bình quân 50. Điểm z là -2 và $+2$ tương ứng với 38 và 62.

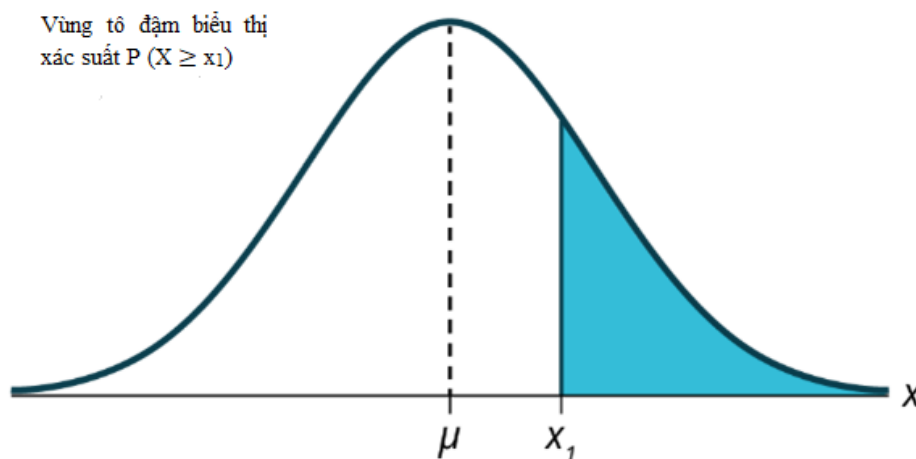
Khoảng 99,7% giá trị x nằm trong ba độ lệch chuẩn so với bình quân. Do đó, khoảng 99,7% giá trị x nằm giữa $-3\sigma = (-3)(6) = -18$ và $3\sigma = (3)(6) = 18$ của giá trị trung bình 50. Các giá trị $50 - 18 = 32$ và $50 + 18 = 68$ nằm trong ba độ lệch chuẩn so với số bình quân 50. Điểm z lần lượt là -3 và $+3$ tương ứng cho 32 và 68.

Bài tập vận dụng 6.2

Giả định X có một phân phối chuẩn với bình quân 25 và độ lệch chuẩn 5,68% của các giá trị x nằm giữa những giá trị?

6.2 Sử dụng phân phối chuẩn

Khu vực được tô đậm trong biểu đồ sau đây biểu thị khu vực bên phải x . Khu vực này được biểu diễn bằng xác suất $P(X > x)$. Trong các giáo trình thống kê thường có các **bảng phân phối chuẩn** (normal table), cung cấp xác suất giữa số bình quân (số không đổi với phân phối chuẩn đơn vị) và một giá trị cụ thể như x_1 . Đó chính là phần không tô đậm của biểu đồ từ số bình quân đến x_1 .



Hình 6.4

Vì phân phối chuẩn là đối xứng, nếu x_1 cùng khoảng cách về bên trái của số bình quân thì phần diện tích, xác suất, ở đuôi bên trái, sẽ giống với diện tích được tô đậm ở đuôi bên phải. Ngoài ra, cần lưu ý rằng vì tính đối xứng của phân phối này, một nửa xác suất nằm ở bên phải của số bình quân μ và một nửa còn lại nằm ở bên trái của số bình quân μ .

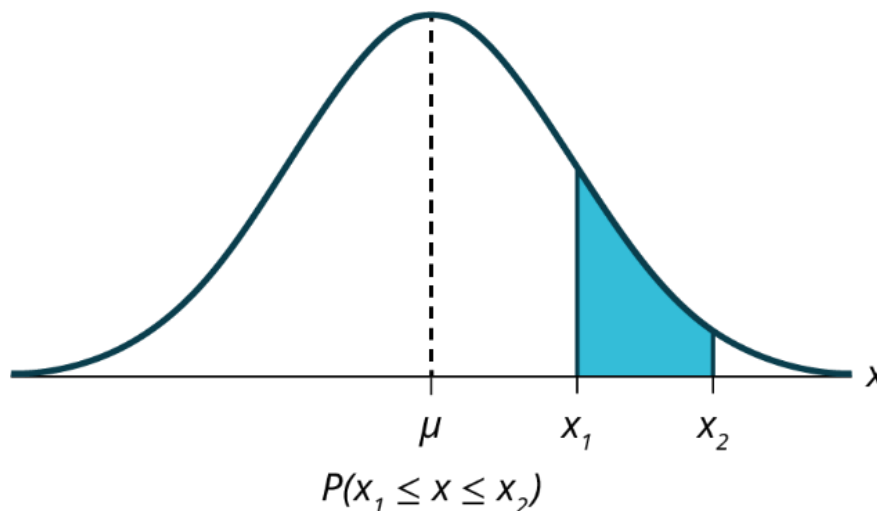
Tính toán xác suất

Để tìm xác suất cho các hàm mật độ xác suất có biến ngẫu nhiên liên tục, chúng ta cần tính diện tích phía dưới đường cong của hàm tại các giá trị x mà chúng ta quan tâm. Đối với phân phối chuẩn, điều này thoạt nhìn có vẻ khó khăn do sự phức tạp của công thức. Tuy nhiên, có một cách đơn giản để có được điều chúng ta cần. Dưới đây là công thức cho phân phối chuẩn:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

Nhìn vào công thức phân phối chuẩn ta không thấy rõ làm cách nào để tính được xác suất theo cách giống như với các hàm xác suất trước đây. Trước đó, ta chỉ việc thay dữ liệu vào công thức và tính toán.

Để giải câu đố này, điều quan trọng là chúng ta hiểu rằng diện tích dưới đồ thị của một hàm mật độ xác suất chính là xác suất.



Hình 6.5

Đồ thị này cho thấy rằng diện tích giữa x_1 và x_2 chính là xác suất như được thể hiện trong công thức: $P(x_1 \leq X \leq x_2)$.

Công cụ toán học cần thiết để tìm diện tích dưới một đường cong là phép tính tích phân. Tích phân của hàm mật độ xác suất chuẩn trong khoảng giữa hai điểm x_1 và x_2 là diện tích dưới đường cong giữa hai điểm này và cũng chính là xác suất giữa hai điểm đó.

Việc tính toán các tích phân này không hề dễ dàng và có thể tốn rất nhiều thời gian. Nhưng bây giờ, ta hãy nhớ rằng có vô số phân phối chuẩn nữa, chúng ta có thể xem xét một phân phối chuẩn có số bình quân bằng 0 và độ lệch chuẩn bằng 1. Phân phối chuẩn đặc biệt này được gọi là Phân phối chuẩn đơn vị. Khi thay các giá trị này vào công thức, nó được rút gọn thành một phương trình rất đơn giản. Giờ đây, ta có thể dễ dàng tính toán tất cả các xác suất cho bất kỳ giá trị x nào trong phân phối chuẩn đặc biệt này, có số bình quân bằng 0 và độ lệch chuẩn bằng 1. Các xác suất tương ứng đã được tính sẵn và có thể tìm thấy ở

phụ lục của giáo trình hoặc trên mạng Internet. Những bảng này được trình bày theo nhiều cách khác nhau. Bảng trong giáo trình này là dạng trình bày phổ biến nhất và được hiển thị xác suất cho một nửa phân phối bắt đầu từ 0 (số bình quân) và mở rộng ra ngoài. Diện tích được tô đậm trong đồ thị ở đầu bảng (trong phần Bảng thống kê) - đại diện cho xác suất từ 0 đến giá trị cụ thể z được ghi trên trục hoành, z .

Vấn đề là dữ liệu của chúng ta nếu tuân theo phân phối chuẩn thì cũng hiếm khi có bình quân bằng 0 và độ lệch chuẩn đúng bằng 1. Giải pháp là chuyển đổi phân phối bình quân μ và độ lệch chuẩn σ sang phân phối chuẩn đơn vị mới này. Phân phối chuẩn đơn vị có một biến ngẫu nhiên được gọi là Z .

Giả sử cần tìm xác suất của z có giá trị một độ lệch chuẩn. Sử dụng bảng phân phối chuẩn đơn vị (thường được gọi là bảng phân phối chuẩn – normal table), ta nhìn cột z , tìm xuống đến hàng 1,0 và sau đó đọc ở cột 0. Giá trị 0,3413 là xác suất để z từ 0 đến 1 độ lệch chuẩn. Ở đầu bảng (trong phần Bảng Thống kê) ta thấy phần diện tích được tô đậm trong đồ thị thể hiện được cung cấp trong bảng. Bảng z giúp ta giải quyết bài toán tích phân, nhưng chỉ khi dữ liệu của chúng ta có số bình quân bằng 0 và độ lệch chuẩn bằng 1.

Tuy nhiên, điểm mấu chốt ở đây là xác suất cho một độ lệch chuẩn trong một phân phối chuẩn thì giống nhau trên mọi phân phối chuẩn. Nếu quần thể dữ liệu có số bình quân là 10 và độ lệch chuẩn là 5, thì xác suất từ 10 đến 15 (một độ lệch chuẩn) giống như xác suất từ 0 đến 1 (một độ lệch chuẩn) trong phân phối chuẩn đơn vị. Để tính xác suất (tức là diện tích dưới đường cong) cho bất kỳ phân phối chuẩn nào, chúng ta chỉ cần chuyển đổi phân phối chuẩn cụ thể sang phân phối chuẩn đơn vị và tra cứu câu trả lời trong bảng Z . Nhắc lại, đây là **công thức chuẩn hóa** (standardizing formula):

$$z = \frac{x - \mu}{\sigma}$$

Trong đó z là giá trị trên phân phối chuẩn đơn vị, x là giá trị từ phân phối chuẩn mà người ta muốn chuyển đổi sang phân phối chuẩn đơn vị, μ và σ lần lượt là bình quân và độ lệch chuẩn của phân phối quần thể đó. Cần nhớ rằng phương trình sử dụng μ và σ là các tham số quần thể. Khi làm việc với xác suất ở đây, chúng ta vẫn luôn làm việc với quần thể có các giá trị tham số đã biết và phân phối đã biết. Điều quan trọng cần lưu ý là vì phân phối chuẩn có tính đối xứng nên không quan trọng là điểm z dương hay âm khi tính xác suất. Một độ lệch chuẩn về bên trái (điểm z âm) vẫn có diện tích (xác suất) tương tự như một độ lệch chuẩn về bên phải (điểm z dương). Đó là lý do tại sao Bảng phân phối chuẩn đơn vị không cung cấp số liệu (diện tích) cho phía bên trái của phân phối. Do tính đối xứng này, công thức chuẩn hóa đôi khi được viết là:

$$z = \frac{|x - \mu|}{\sigma}$$

Trong đó các đường thẳng đứng $||$ trong công thức có nghĩa là giá trị tuyệt đối của số.

Công thức chuẩn hóa thực tế đang tính toán xem giá trị x cách số bình quân bao nhiêu độ lệch chuẩn trong chính phân phối của nó. Công thức chuẩn hóa và khái niệm đếm các độ lệch chuẩn từ số bình quân là bí quyết của tất cả những gì chúng ta sẽ làm trong khoá học thống kê này. Đó là vì tất cả thống kê đều quy về khái niệm biến thiên, và việc đếm các độ lệch chuẩn chính là cách đo lường sự biến thiên đó.

Công thức này, dưới nhiều hình thức khác nhau, sẽ xuất hiện lặp đi lặp lại trong giáo trình này.

Ví dụ 6.3

Điểm thi cuối kỳ trong một lớp thống kê được tuân theo phân phối chuẩn với bình quân là 63 và độ lệch chuẩn là 5.

Vấn đề

- Tìm xác suất để một sinh viên được chọn ngẫu nhiên đạt điểm cao hơn 65 trong kỳ thi.
- Tìm xác suất để một sinh viên được chọn ngẫu nhiên đạt điểm thấp hơn 85.

Lời giải

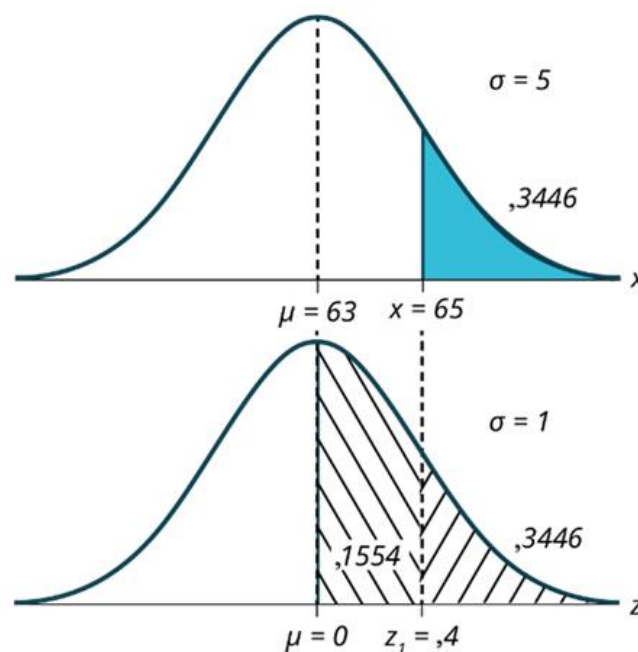
a.

Gọi X = điểm số trong bài thi cuối kỳ. $X \sim N(63, 5)$, trong đó $\mu = 63$ và $\sigma = 5$.

Vẽ đồ thị.

Sau đó, tìm $P(X > 65)$.

$$P(X > 65) = 0,3446$$



Hình 6.6

$$z_1 = \frac{x_1 - \mu}{\sigma} = \frac{65 - 63}{5} = 0,4$$

$$P(X \geq x_1) = P(Z \geq z_1) = 0,3446$$

Xác suất để một sinh viên được chọn ngẫu nhiên đạt điểm cao hơn 65 là 0,3446. Đây là cách chúng ta tìm ra kết quả này.

Bảng phân phối chuẩn cung cấp xác suất từ 0 đến một giá trị z_1 . Đối với bài toán này, câu hỏi có thể được viết lại như sau: $P(X \geq 65) = P(Z \geq z_1)$, đây là xác suất nằm ở phần đuôi bên phải của phân phối. Để tìm diện tích này, ta áp dụng công thức sẽ là: $0,5 - P(X \leq 65)$. Một nửa xác suất nằm phía trên số bình quân bởi vì phân phối chuẩn là đối xứng. Đồ thị trên cho thấy diện tích ở phần đuôi được xác định bằng cách trừ phần đó từ số bình quân, 0, đến giá trị z_1 . Câu trả lời cuối cùng là: $P(X \geq 65) = P(Z \geq 0,4) = 0,3446$.

$$z_1 = \frac{65-63}{5} = 0,4$$

Diện tích phần bên trái từ z_1 đến bình quân (số không) là 0,1554

Do đó $P(x > 65) = P(z > 0,4) = 0,5 - 0,1554 = 0,3446$.

b.

Một điểm số là 85, cách trung bình 63 đến 22 điểm, tức là:

$$z = \frac{x - \mu}{\sigma} = \frac{85 - 63}{5} = 4,4$$

giá trị này lớn hơn giá trị tối đa trong Bảng phân phối chuẩn đơn vị. Do đó, xác suất để một sinh viên đạt điểm thấp hơn 85 là xấp xỉ một (hoặc 100%).

Điểm 85 có 4,4 độ lệch chuẩn cách xa bình quân 63, vượt quá phạm vi của bảng phân phối chuẩn đơn vị. Do đó, xác suất để một sinh viên đạt điểm thấp hơn 85 là xấp xỉ một (hoặc 100%).

Bài tập vận dụng 6.3

Điểm đánh gôn của một đội tuyển trường học tuân theo phân phối chuẩn với bình quân là 68 và độ lệch chuẩn là 3. Tìm xác suất để một vận động viên đánh gôn được chọn ngẫu nhiên có điểm thấp hơn 65.

Ví dụ 6.4

Một máy tính cá nhân được sử dụng tại nhà cho nhiều mục đích như: công việc, nghiên cứu, giao tiếp, tài chính cá nhân, giáo dục, giải trí, mạng xã hội và vô số những hoạt động khác. Giả sử số giờ trung bình mà một máy tính cá nhân gia đình được sử dụng cho mục đích giải trí là hai giờ mỗi ngày. Giả định thời gian sử dụng cho giải trí là được tuân theo phân phối chuẩn và độ lệch chuẩn của thời gian là nửa giờ.

Vấn đề

a. Tìm xác suất máy tính cá nhân gia đình được sử dụng để giải trí trong khoảng từ 1,8 đến 2,75 giờ mỗi ngày.

b. Tìm số giờ tối đa mỗi ngày sử dụng máy tính cá nhân để giải trí của nhóm gia đình trong tứ phân vị nhỏ nhất.

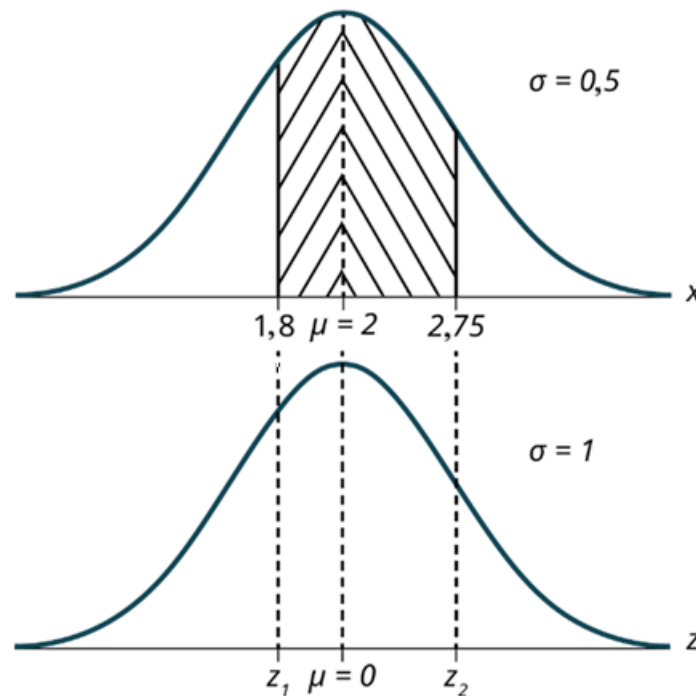
Lời giải

a.

Gọi X = lượng thời gian (tính bằng giờ) một máy tính cá nhân gia đình được sử dụng cho mục đích giải trí. $X \sim N(2, 0,5)$ trong đó $\mu = 2$ và $\sigma = 0,5$.

Tìm $P(1,8 < X < 2,75)$.

Xác suất ta đang tìm kiếm là diện tích **giữa** $x = 1,8$ và $x = 2,75$. $P(1,8 < X < 2,75) = 0,5886$.



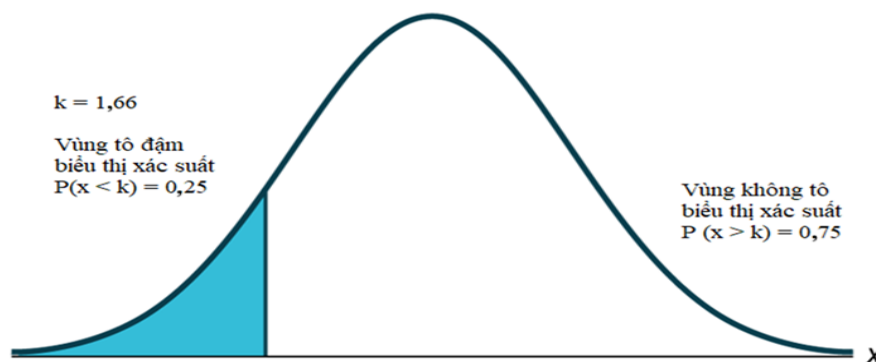
Hình 6.7

$$P(1,8 \leq x \leq 2,75) = P(z_1 \leq z \leq z_2)$$

Xác suất một máy tính cá nhân gia đình được sử dụng từ 1,8 đến 2,75 giờ mỗi ngày để giải trí là 0,5886.

b.

Để tìm số giờ tối đa mỗi ngày sử dụng máy tính cá nhân để giải trí của các gia đình trong nhóm tứ phân vị nhỏ nhất, hãy **tìm phân vị thứ 25**, ký hiệu k sao cho $P(X < k) = 0,25$



Hình 6.8

$$f(z) = 0,5 - 0,25 = 0,25,$$

do đó $z \approx -0,675$ (hoặc chỉ 0,67 khi sử dụng bảng)

$$z = \frac{x - \mu}{\sigma} = \frac{x - 2}{0,5} = -0,675,$$

vì vậy $x = -0,675 * 0,5 + 2 = 1,66$ giờ.

Số giờ tối đa mỗi ngày sử dụng máy tính cá nhân để giải trí của các gia đình trong nhóm tứ phân vị nhỏ nhất là 1,66 giờ.

Bài tập vận dụng 6.4

Điểm số chơi đánh gôn của một đội trường học tuân theo phân phối chuẩn với số bình quân là 68 và độ lệch chuẩn là 3. Tìm xác suất một người chơi đánh gôn đạt điểm trong khoảng từ 66 đến 70.

Ví dụ 6.5

Tại Hoa Kỳ, độ tuổi từ 13 đến 55+ của người dùng điện thoại thông minh gần như tuân theo phân phối chuẩn với bình quân và độ lệch chuẩn xấp xỉ lần lượt là 36,9 năm và 13,9 năm.

Vấn đề

- Xác định xác suất một người dùng điện thoại thông minh ngẫu nhiên trong độ tuổi từ 13 đến 55+ là có tuổi nằm trong khoảng từ 23 đến 64,7 tuổi.
- Xác định xác suất một người dùng điện thoại thông minh được chọn ngẫu nhiên trong độ tuổi từ 13 đến 55+ có tuổi không vượt quá 50,8.

Lời giải

- 0,8186
- 0,8413

Bài tập vận dụng 6.5

Sử dụng thông tin trong Ví dụ 6.5 để trả lời các câu hỏi sau.

- Tìm số phân vị thứ 30 và diễn giải nó trong một câu hoàn chỉnh.
- Tìm xác suất người dùng điện thoại thông minh được chọn ngẫu nhiên trong khoảng độ tuổi từ 13 đến 55+ có tuổi nhỏ hơn 27 tuổi.

Ví dụ 6.6

Một người nông dân trồng quýt nhận thấy rằng đường kính của những quả quýt được thu hoạch trên trang trại của họ tuân theo phân phối chuẩn với bình quân $\mu = 5,85$ cm và độ lệch chuẩn là $\sigma = 0,24$ cm.

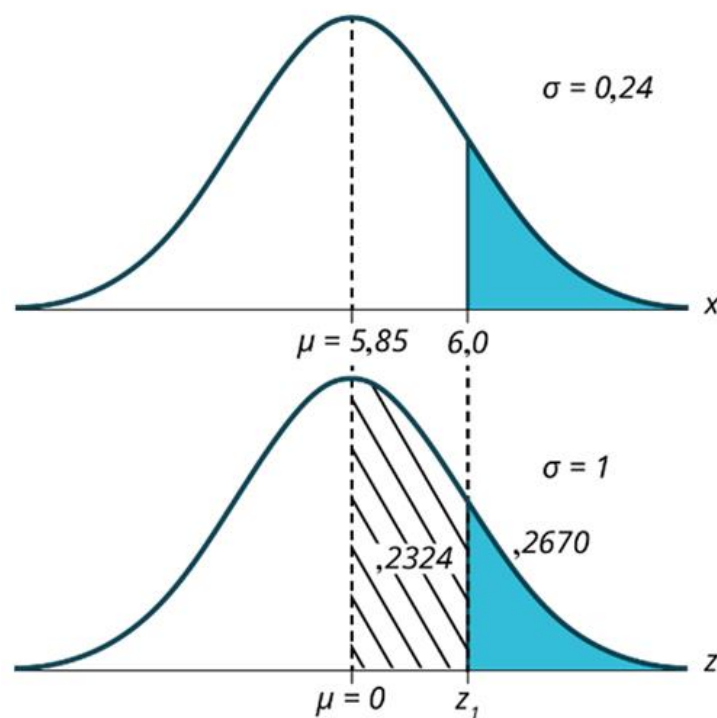
Vấn đề

Tìm xác suất một quả quýt được chọn ngẫu nhiên từ trang trại này có đường kính lớn hơn 6,0 cm. Phác thảo đồ thị.

20% giữa số quýt của trang trại này có đường kính từ ____ đến ____.

Lời giải

a.



Hình 6.9

$$z_1 = \frac{6-5,85}{0,24} = 0,625$$

$$P(x \geq 6) = P(z \geq 0,625) = 0,2670$$

b.

$$f(z) = \frac{0,20}{2} = 0,10, \text{ vì vậy } z \approx \pm 0,25$$

$$z = \frac{x - \mu}{\sigma} = \frac{x - 5,85}{0,24} = \pm 0,25 \rightarrow \pm 0,25 * 0,24 + 5,85 = (5,79; 5,91)$$

Bài tập vận dụng 6.6

Sử dụng thông tin từ Ví dụ 6.6, hãy trả lời các câu hỏi sau:

- 40% giữa số quýt của trang trại này nằm trong khoảng từ ____ đến ____.
- Tìm số phân vị thứ 16 và diễn giải trong một câu hoàn chỉnh.

LƯU Ý

Sử dụng Excel, R để tính phân phối chuẩn.

Excel: xác suất phân phối chuẩn $P(X < x)$ hàm NORM.DIST(x, m, s, 1); m là bình quân, s là độ lệch chuẩn của biến X. Để tìm xác suất phân phối chuẩn đơn vị $P(Z < z)$, ta có thể sử dụng hàm NORM.DIST với m = 0, s = 1; hoặc sử dụng hàm NORM.S.DIST(z, 1).

Nếu biết xác suất $p = P(X < x)$, ta có thể tìm giá trị x bằng hàm NORM.INV(p, m, s); m là bình quân, s là độ lệch chuẩn của biến X. Tương tự có thể tìm giá trị z sao cho $p = P(Z < z)$ bằng hàm NORM.S.INV(p).

R/stats package: xác suất phân phối chuẩn $P(X < x)$ hàm pnorm(x, m, s, 1); m là bình quân, s là độ lệch chuẩn của biến X. Để tìm xác suất phân phối chuẩn đơn vị $P(Z < z)$, ta có thể sử dụng hàm pnormr với m = 0, s = 1.

Nếu biết xác suất $p = P(X < x)$, ta có thể tìm giá trị x bằng hàm qnorm(p, m, s); m là bình quân, s là độ lệch chuẩn của biến X.

6.3 Ước lượng phân phối nhị thức bằng phân phối chuẩn

Chúng ta đã tìm hiểu trước đó rằng các hàm mật độ xác suất khác nhau có thể là phân phối giới hạn của những hàm khác; do đó, chúng ta có thể ước tính một phân phối bằng một phân phối khác trong một số trường hợp nhất định. Ở đây, chúng ta sẽ thấy rằng phân phối chuẩn có thể được sử dụng để ước tính một quá trình nhị thức. Ở chương trước, phân phối Poisson đã được dùng để ước tính nhị thức, và nhị thức đã được sử dụng để ước tính phân phối siêu bội.

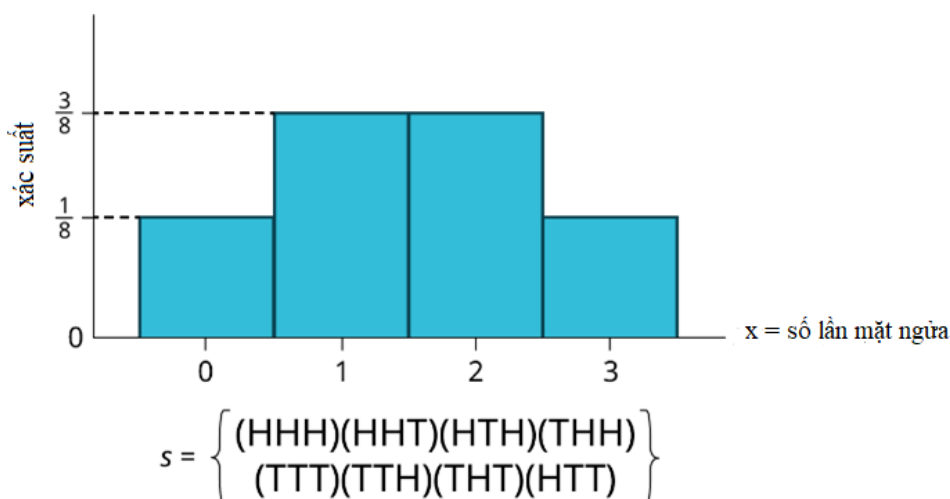
Trong trường hợp mối quan hệ giữa phân phối siêu bội và nhị thức, chúng ta thấy một quá trình nhị thức giả định rằng xác suất thành công là không đổi qua các lần thử khác

nhau: kết quả đồng xu ra mặt ngửa ở lần tung trước không hề ảnh hưởng đến xác suất ra mặt ngửa ở lần tung tiếp theo. Trong phân phối siêu bội, đây là cốt lõi của vấn đề vì thực nghiệm siêu bội giả định rằng bất kỳ "lần rút" nào cũng là không hoàn lại. Nếu rút không hoàn lại, thì tất cả các "lần rút" tiếp theo đều là xác suất có điều kiện. Chúng ta cũng đã thấy rằng nếu thực nghiệm siêu bội chỉ rút một tỷ lệ nhỏ trong tổng số các phần tử (trong quần thể), thì chúng ta có thể bỏ qua tác động lên xác suất qua các lần rút khác nhau.

Giả sử có 312 lá bài trong một bộ bài bao gồm 6 bộ bài tây thông thường. Nếu thực nghiệm yêu cầu rút chỉ 10 lá bài, ít hơn 5% của tổng số, thì chúng ta có thể chấp nhận ước lượng phân phối này bằng phân phối nhị thức, mặc dù đây về bản chất là phân phối siêu bội vì các lá bài được rút mà không hoàn lại.

Tương tự, phân phối Poisson được coi là một ước lượng thích hợp của phân phối nhị thức trong một số trường hợp nhất định. Trong chương biến ngẫu nhiên rời rạc, chúng ta đã thấy rằng nếu số lần thử lớn và xác suất thành công nhỏ thì Poisson có thể ước lượng phân phối nhị thức với những kết quả tốt. Cần lưu ý một lần nữa là những quy tắc này không khẳng định là ước lượng xác suất sẽ hoàn toàn chính xác, mà chỉ có nghĩa là sự chênh lệch sẽ ở hàng chữ số thập phân thứ ba hoặc thứ tư và do đó có thể bỏ qua.

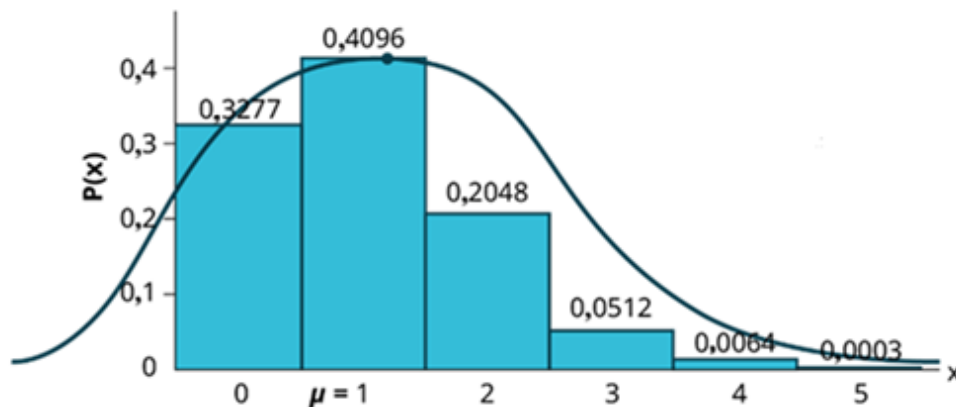
Ở đây, một lần nữa, chúng ta sẽ thấy phân phối chuẩn đưa ra các ước lượng đặc biệt chính xác về một quá trình nhị thức trong một số trường hợp nhất định. Hình 6.10 là phân phối tần suất của một quá trình nhị thức cho thực nghiệm tung ba đồng xu, trong đó biến ngẫu nhiên là số lần ra mặt ngửa. Không gian mẫu được liệt kê bên dưới phân phối (H = ngửa, T = sấp). Thực nghiệm giả định rằng xác suất thành công là 0,5; do đó, xác suất thất bại, mặt sấp, cũng là 0,5. Khi quan sát Hình 6.10, chúng ta có thể thấy là phân phối có tính đối xứng. Gốc rễ của kết quả này xuất phát từ xác suất thành công và thất bại đều bằng nhau, 0,5. Nếu xác suất thành công nhỏ hơn 0,5, phân phối sẽ bị lệch phải. Thật vậy, khi xác suất thành công giảm, mức độ lệch tăng lên. Nếu xác suất thành công tăng từ 0,5, thì độ lệch tăng ở bên trái, dẫn đến phân phối lệch trái.



Hình 6.10

Độ lệch của phân phối nhị thức có ý nghĩa quan trọng bởi vì nếu ta muốn ước lượng nó bằng phân phối chuẩn, chúng ta cần nhận thức rằng phân phối chuẩn là đối xứng. Phân

phối nhị thức cơ bản càng gần với sự đối xứng, thì ước lượng bởi phân phối chuẩn càng chính xác. Hình 6.11 minh họa một phân phối chuẩn đối xứng được vẽ chồng lên đồ thị của phân phối nhị thức với $p=0,2$ và $n = 5$. Có thể thấy rõ sự sai lệch giữa xác suất ước lượng bởi phân phối chuẩn và xác suất thật sự của phân phối nhị thức. Do đó, để có thể sử dụng phân phối chuẩn cho ước lượng phân phối nhị thức yêu cầu np VÀ $n(1 - p)$ đều lớn hơn năm. Một lần nữa, đây là một quy tắc kinh nghiệm, nhưng nó hiệu quả và mang đến các ước tính chấp nhận được về xác suất nhị thức.



Số lần mặt ngửa

Hình 6.11

Ví dụ 6.7

Người ta biết rằng chỉ có 10% chó con giống chăn cừu Úc được sinh ra với sự "đối xứng hoàn hảo" trong ba màu lông: đen, trắng và đồng. "Đối xứng hoàn hảo" được định nghĩa là các mảng màu được phân bố đều nhau ở cả hai phía của cơ thể chó khi nhìn chính diện và chia theo đường trung tâm mặt. Một trại chó sẽ có được đánh giá là có danh tiếng tốt về việc nhân giống chó chăn cừu Úc nếu họ có tỷ lệ chó đáp ứng được tiêu chí đặc điểm đối xứng này cao. Trong 5 năm qua, trong số 100 con chó được sinh ra tại Dundee Kennels, 16 con được sinh ra với đặc điểm phối màu lông này.

Vấn đề

Tìm xác suất để trong 100 lần sinh, có hơn 16 con chó có đặc điểm này?

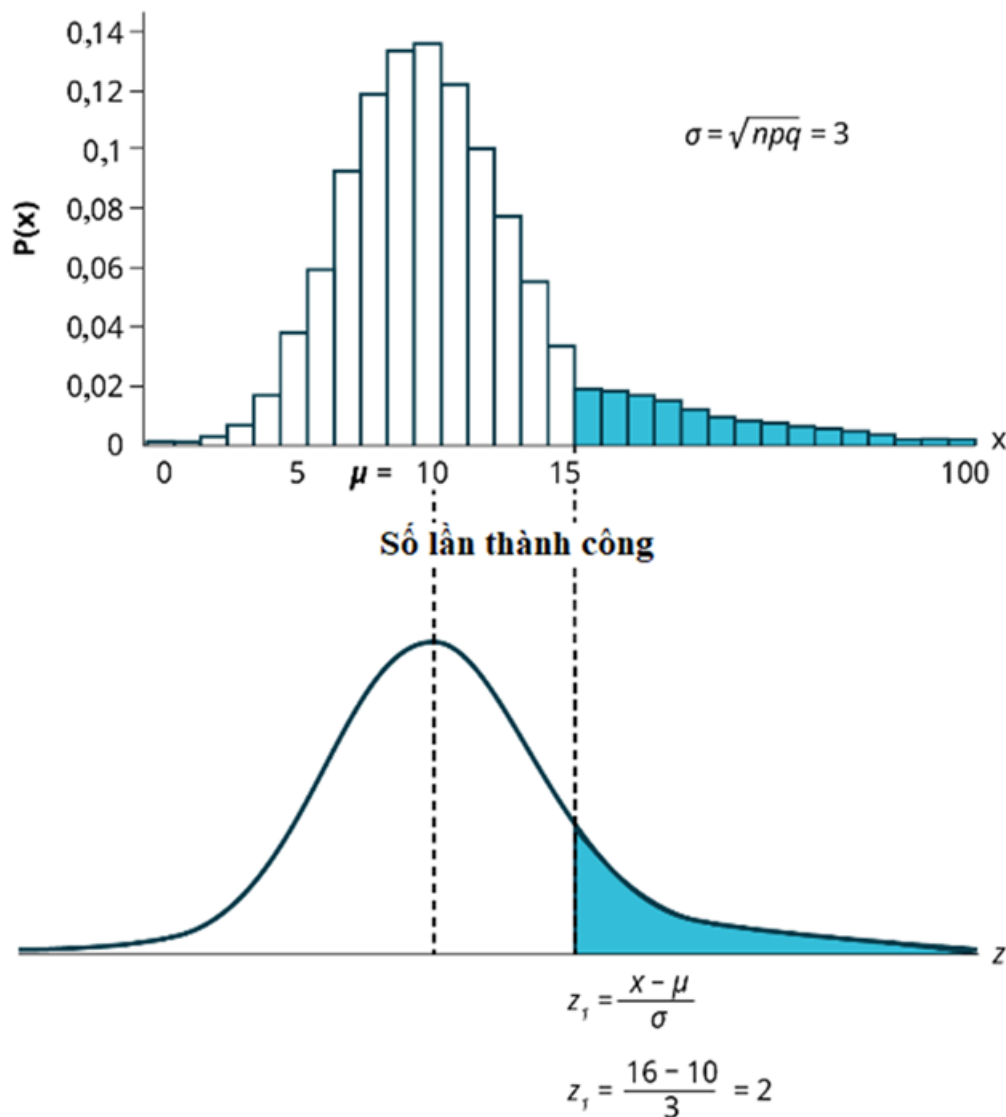
Lời giải

Nếu chúng ta giả định rằng màu lông của một con chó độc lập với màu lông của những con chó khác (một giả định có phần táo bạo), thì đây trở thành một bài toán xác suất nhị thức kinh điển.

Phát biểu về xác suất được yêu cầu là $1 - [p(x=0) + p(x=1) + p(x=2) + \dots + p(x=16)]$. Điều này đòi hỏi chúng ta phải tính toán 17 công thức nhị thức và cộng chúng lại với nhau, sau đó lấy 1 trừ đi kết quả tổng để tìm xác suất bên phải của phân phối. Ngoài ra, chúng ta có thể sử dụng phân phối chuẩn để có được câu trả lời một cách nhanh chóng chấp nhận được.

Đầu tiên, chúng ta cần kiểm tra xem phân phối nhị thức có đủ đối xứng để sử dụng phân phối chuẩn hay không. Chúng ta biết rằng nhị thức cho bài toán này bị lệch vì xác suất thành công, 0,1, không giống với xác suất thất bại, 0,9. Tuy nhiên, cả hai $np = 10$ và $n(1 - p) = 90$ đều lớn hơn 5, ngưỡng để sử dụng phân phối chuẩn để ước tính phân phối nhị thức.

Hình 6.12 dưới đây cho thấy phân phối nhị thức và đánh dấu khu vực chúng ta muốn biết. Bình quân của nhị thức là 10, cũng được đánh dấu, và độ lệch chuẩn được viết ở bên cạnh đồ thị: $\sigma = \sqrt{npq} = 3$. Diện tích dưới phân phối từ 0 đến 16 là xác suất được yêu cầu và đã được tô đậm. Bên dưới phân phối nhị thức là một phân phối chuẩn được sử dụng để ước tính xác suất này. Xác suất đó cũng đã được tô đậm.



Hình 6.12

Ước lượng phân phối nhị thức bằng phân phối chuẩn cũng cần chuẩn hóa như đã thực hiện trước đây. Để tính xác suất từ 16 đến vô cực dương (hoặc là 100 trong trường hợp này), ta cần xác định bao nhiêu độ lệch chuẩn mà giá trị 16 cách xa bình quân là 10.

$$z = \frac{x - \mu}{\sigma} = \frac{16 - 10}{3} = 2$$

Chúng ta đang tìm xác suất vượt quá hai độ lệch chuẩn, một sự kiện rất hiếm xảy ra trong phân phối chuẩn. Chúng ta tra cứu hai độ lệch chuẩn trong bảng phân phối chuẩn đơn vị và tìm thấy diện tích từ 0 đến hai độ lệch chuẩn là 0,4772. Tuy nhiên, chúng ta quan tâm đến phần đuôi bên phải, vì vậy chúng ta trừ 0,4772 khỏi 0,5 và do đó tìm thấy diện tích ở phần đuôi. Kết luận của chúng ta là xác suất một chuồng chó có 16 con chó có "tính đối xứng hoàn hảo" là 0,0228. Điều này cho thấy Dundee Kennels có một thành tích xuất sắc về vấn đề này.

Về mặt toán học, chúng ta viết công thức như sau:

$$1 - [p(x = 0) + p(x = 1) + p(x = 2) + \dots + p(x = 16)] = p(x > 16) = p(z > 2) = 0,0228.$$

Bài tập vận dụng 6.7

Giả sử rằng chỉ có 10% đồ nội thất bằng gỗ do một thợ mộc làm ra được làm theo đúng kích thước chính xác trong bản thiết kế. Trong 3 năm qua, người thợ mộc đã làm 100 món đồ nội thất bằng gỗ và đã tìm thấy trong đó có 16 món được làm theo đúng kích thước trong bản thiết kế.

Tìm xác suất trong 100 món đồ nội thất bằng gỗ, có hơn 16 món được làm theo đúng kích thước của bản thiết kế?

Thuật ngữ chính

Điểm chuẩn hóa z (z-score) là phép biến đổi tuyến tính có dạng $z = \frac{x - \mu}{\sigma}$ hoặc được viết là $z = \frac{|x - \mu|}{\sigma}$; nếu phép biến đổi này được áp dụng cho bất kỳ phân phối chuẩn nào $X \sim N(\mu, \sigma)$ thì kết quả là phân phối chuẩn đơn vị $Z \sim N(0,1)$. Nếu phép biến đổi này được áp dụng cho bất kỳ giá trị cụ thể x nào của biến ngẫu nhiên có số bình quân μ và độ lệch chuẩn σ , thì kết quả được gọi là điểm chuẩn hóa z của x . Điểm chuẩn hóa z cho phép chúng ta so sánh dữ liệu có cùng phân phối chuẩn nhưng có thang đo khác nhau. Điểm chuẩn hóa z là số độ lệch chuẩn của một giá trị x cụ thể so với bình quân của nó.

Phân phối chuẩn (normal distribution) là biến ngẫu nhiên liên tục (RV) có pdf $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, trong đó μ là bình quân của phân phối và σ là độ lệch chuẩn; ký hiệu: $x \sim N(\mu, \sigma)$. Nếu $\mu = 0$ và $\sigma = 1$, biến ngẫu nhiên Z , được gọi là phân phối chuẩn đơn vị.

Phân phối chuẩn đơn vị (Standard normal distribution) là biến ngẫu nhiên liên tục (RV) $X \sim N(0, 1)$; khi X tuân theo phân phối chuẩn đơn vị, nó thường được ký hiệu là $Z \sim N(0, 1)$.

Tóm tắt chương

6.1 Phân phối chuẩn đơn vị

Điểm chuẩn hóa z (z-score) là một giá trị đã được chuẩn hóa. Phân phối của nó là phân phối chuẩn, $Z \sim N(0, 1)$. Bình quân của Z là 0 và độ lệch chuẩn là 1. Nếu z là điểm chuẩn hóa cho một giá trị x từ phân phối chuẩn $N(\mu, \sigma)$ thì z chỉ cho chúng ta biết x cách bình quân μ bao nhiêu độ lệch chuẩn phía trên (bên phải, lớn hơn) hoặc phía dưới (bên trái, nhỏ hơn).

6.3 Ước lượng phân phối nhị thức bằng phân phối chuẩn

Phân phối chuẩn (normal distribution), một phân phối liên tục, là phân phối xác suất quan trọng nhất. Đồ thị của nó có dạng hình chuông. Đường cong hình chuông này được sử dụng trong hầu hết mọi lĩnh vực khoa học. Do nó là phân phối liên tục nên tổng diện tích dưới đường cong là một. Các tham số của phân phối chuẩn là bình quân μ và độ lệch chuẩn σ . Một phân phối chuẩn đặc biệt, được gọi là phân phối chuẩn đơn vị là phân phối của các điểm chuẩn hóa z . Bình quân của nó là 0 và độ lệch chuẩn của nó là 1.

Tóm tắt công thức

Giới thiệu

$$x \sim N(\mu, \sigma)$$

μ = bình quân; σ = độ lệch chuẩn

6.1 Phân phối chuẩn đơn vị

$$Z \sim N(0, 1)$$

z = giá trị chuẩn hóa (điểm số z)

bình quân = 0; độ lệch chuẩn = 1

Để tìm giá trị quan sát được x , khi biết được điểm số z :

$$x = \mu + (z)\sigma$$

$$\text{Điểm chuẩn hóa } Z : z = \frac{x - \mu}{\sigma} \text{ or } z = \frac{|x - \mu|}{\sigma}$$

Z = biến ngẫu nhiên cho điểm chuẩn hóa z

$$Z \sim N(0, 1)$$

6.3 Ước lượng phân phối nhị thức bằng phân phối chuẩn

Phân phối chuẩn: $X \sim N(\mu, \sigma)$ trong đó μ là bình quân và σ là độ lệch chuẩn.

Phân phối chuẩn đơn vị: $Z \sim N(0, 1)$

Tài liệu tham khảo

6.1 Phân phối chuẩn đơn vị

“Blood Pressure of Males and Females.” StatCrunch, 2013. Available online at <http://www.statcrunch.com/5.0/viewreport.php?reportid=11960> (accessed May 14, 2013).

“The Use of Epidemiological Tools in Conflict-affected populations: Open-access educational resources for policy-makers: Calculation of z-scores.” London School of Hygiene and Tropical Medicine, 2009. Available online at http://conflict.lshtm.ac.uk/page_125.htm (accessed May 14, 2013).

“2012 College-Bound Seniors Total Group Profile Report.” CollegeBoard, 2012. Available online at <http://media.collegeboard.com/digitalServices/pdf/research/TotalGroup-2012.pdf> (accessed May 14, 2013).

“Digest of Education Statistics: ACT score average and standard deviations by sex and race/ethnicity and percentage of ACT test takers, by selected composite score ranges and planned fields of study: Selected years, 1995 through 2009.” National Center for Education Statistics. Available online at http://nces.ed.gov/programs/digest/d09/tables/dt09_147.asp (accessed May 14, 2013).

Data from the San Jose Mercury News.

Data from The World Almanac and Book of Facts.

“List of stadiums by capacity.” Wikipedia. Available online at https://en.wikipedia.org/wiki/List_of_stadiums_by_capacity (accessed May 14, 2013).

Data from the National Basketball Association. Available online at www.nba.com (accessed May 14, 2013).

6.2 Sử dụng phân phối chuẩn

“Naegle’s rule.” Wikipedia. Available online at http://en.wikipedia.org/wiki/Naegle's_rule (accessed May 14, 2013).

“403: NUMMI.” Chicago Public Media & Ira Glass, 2013. Available online at <http://www.thisamericanlife.org/radioarchives/episode/403/nummi> (accessed May 14, 2013).

“Scratch-Off Lottery Ticket Playing Tips.” WinAtTheLottery.com, 2013. Available online at <http://www.winatthelottery.com/public/departments40.cfm> (accessed May 14, 2013).

“Smart Phone Users, By The Numbers.” Visual.ly, 2013. Available online at <http://visual.ly/smart-phone-users-numbers> (accessed May 14, 2013).

“Facebook Statistics.” Statistics Brain. Available online at <http://www.statisticbrain.com/facebook-statistics/> (accessed May 14, 2013).