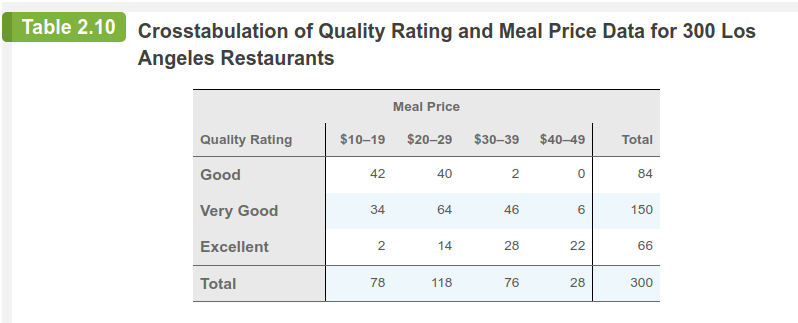
**BẢNG CHÉO**

**Bảng chéo** là một bảng tóm tắt dữ liệu cho hai biến. Mặc dù cả 2 biến có thể là định tính hoặc định lượng, bảng tính có 1 biến là định lượng và biến còn lại là định tính là điều bình thường.

Xem xét ứng dụng sau đây dựa trên dữ liệu từ nhà hàng Zagat. Dữ liệu đánh giá chất lượng và giá bữa ăn của một mãu được thu thập từ 300 nhà hàng nằm trong khu vực Los Angleles.

|  |  |  |
| --- | --- | --- |
| **Restaurant** | **Quality Rating** | **Meal Price ($)** |
| 1 | Good | 18 |
| 2 | Very Good | 22 |
| 3 | Good | 28 |
| 4 | Excellent | 38 |
| 5 | Very Good | 33 |
| 6 | Good | 28 |
| 7 | Very Good | 19 |
| 8 | Very Good | 11 |
| 9 | Very Good | 23 |
| 10 | Good | 13 |

Dữ liệu đánh giá chất lượng của một nhà hàng và giá cả bữa ăn. Đánh giá chất lượng là một biến định tính với các loại đánh giá là tốt, rất tốt, và tuyệt vời. Giá bữa ăn là một biến định lượng khoảng từ 10 đến 49 USD.

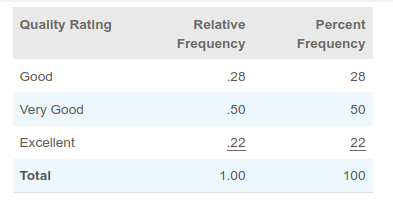


Có 2 nhóm chính: Nhóm đánh giá chất lượng (tốt, rất tốt, xuất sắc), nhóm biến giá (giá bữa ăn).

Ví dụ, nhà hàng 5 được xác định là có một đánh giá chất lượng rất tốt và giá cả bữa ăn là 33 USD. Nhà nhàng này thuộc về ô ở dòng 2 và cột 3 của bảng 2.10. Trong việc xây dựng bảng chéo, đơn giản là đếm số lượng các nhà hàng thuộc về mỗi ô trong bảng chéo.

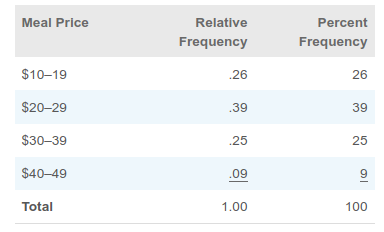
Trong bảng 2.10, có 64 nhà hàng trong mẫu được đánh giá rất tốt và có giá bữa ăn khoảng 20 – 29 USD.

Đem chia các giá trị tổng ở lề bên phải cho bảng chéo cho tổng ta sẽ có được phân phối tần suất và phần trăm của biến đánh giá chất lượng.



Từ phân phối tần suất phần trăm ta thấy có 28% các nhà hàng được đánh giá tốt, 50% được đánh giá rất tốt, và 22% được đánh giá tuyệt vời.

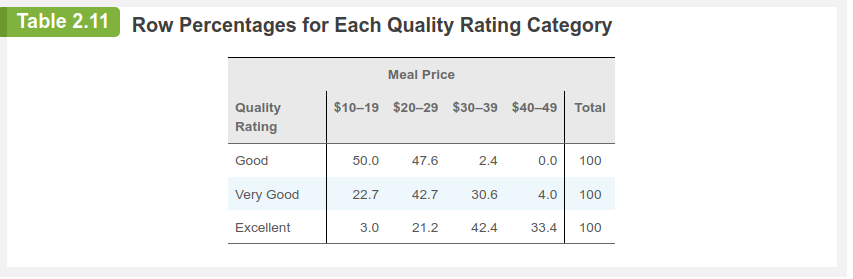
Đem chia các tổng số ở hàng dưới cùng của bảng chéo cho tổng của hàng này, ta sẽ có một phân phối tần suất và phần trăm của biến giá.



Lưu ý rằng cộng các giá trị trong mỗi cột không bằng đúng giá trị tổng, bởi các giá trị tóm tắt được làm tròn. Từ phân phối tần suất phần trăm chúng ta thấy rằng 26% giá bữa ăn đang ở trong nhóm giá thấp nhất (10 – 19 USD), 39% là trong nhóm cao hơn tiếp theo...

Bảng chéo không cho thấy có mỗi liên hệ của các nhóm một cách rõ ràng. Giá trị chính của một bảng chéo cho ta thấy cái nhìn sâu hơn về mỗi liên hệ giữa các biến. Bảng 2.10 cho thấy giá bữa ăn cao có liên quan đến các nhà hàng chất lượng cao và giá thấp hơn liên quan đến các nhà hàng có chất lượng thấp hơn.

Chuyển đổi các con số trong bảng chéo thành tỷ lệ phần trăm theo hàng hoặc theo cột có thể cung cấp cái nhín sâu hơn về các mỗi quan hệ giữa hai biến. Đối với tỷ lệ phần trăm hàng, kết quả chia mỗi tần số trong bảng 2.10 cho tổng hàng tương ứng của nó được thể hiện trong bảng 2.11.



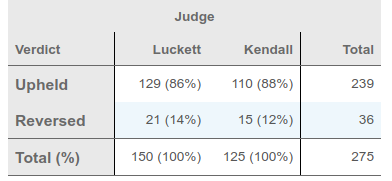
Bảng 2.11 là một phân phối tần suất phần trăm của giá bữa ăn tương ứng với loại đánh giá chất lượng. Các nhà hàng với những đánh giá chất lượng thấp nhất (tốt), thấy rằng tỷ lệ phần trăm lớn nhất là nhà hàng rẻ (50% giá 10 – 19 USD và 47.6% có giá 20 – 29 USD). Các nhà hàng được đánh giá chất lượng cao nhất (tuyệt vời), chúng ta thấy rằng tỷ lệ phần trăm lớn nhất là các nhà hàng đắt tiền (42.4 % giá 30 – 39 USD và 33.4% có 40 – 49 USD). Do đó, có thể nhận định rằng các bữa ăn đắt tiền hơn có liên quan đến các nhà hàng có chất lượng cao hơn.

**Nghịch lý Simpson**

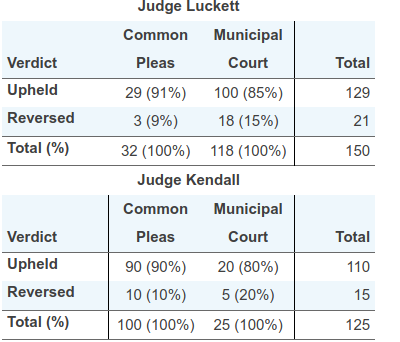
Trong một số trường hợp, các kết luận dựa trên bảng chéo tổng hợp có thể hoàn toàn đảo ngược nếu chúng ta nhìn vào các dữ liệu chưa tổng hợp, gọi là nghịch lý Simpson. Xem xét một ví dụ liên quan đến việc phân tích các bản án của 2 thẩm phán tại 2 loại tòa án.

Thẩm phấn Ron Luckket và Dennis Kendall chủ trì các vụ án ở tòa án Common Pleas và Municipal Courl trong 3 năm qua. Một số bản án họ tuyên bị kháng cáo. Trong hầu hết các trường hợp, tòa phúc thẩm đã giữ nguyên phán quyết ban đầu, nhưng trong một số trường hợp bản án đã bị đảo ngược. Đối với một thẩm phán một bảng chéo đã được xây dựng dựa trên 2 yếu tố: Bản án (giữ nguyên hoặc đảo ngược) và loại Tòa án (Common Pleas và Munucipal). Giả sử rằng 2 bảng chéo sau đó được kết hợp bằng cách tập hợp dữ liệu từ hai toà án. Kết quả bảng chéo tổng hợp có chưa 2 biến: Bản án (giữ nguyên hoặc đảo ngược) và Thẩm phán (Luckett hoặc Kendall). Bảng chéo cho thấy số lượng kháng cáo trong đó bản án được giữ nguyên và một trong số đó đã bị đảo ngược phán quyết cho cả 2 thẩm phán. Lập bảng chéo sau đây cho thấy những kết quả này cùng với tỷ lệ phần trăm cột trong dấu ngoặc đơn bên cạnh mỗi giá trị.

Tỷ lệ phần trăm theo cột cho thấy rằng 14% các bản án của Thầm phán Luckett đã bị đảo ngược nhưng chỉ có 12% các bản án của thẩm phán Kendall đã bị đảo ngược. Do đó, chúng ta có thể kết luận rằng Thẩm phán Kendall đang làm việc hiệu quả hơn bởi vì tỷ lệ phần trăm các bản án của ông được giữ nguyên cao hơn. Tuy vậy có một vấn đề phát sinh với kết luận này.



Các bảng chéo sau đây cho thấy các trường hợp cố gắng của Luckett và Kendall trong 2 tòa án, tỷ lệ phần trăm cũng được hiển thị trong dấu ngoặc đơn bên cạnh mỗi giá trị.



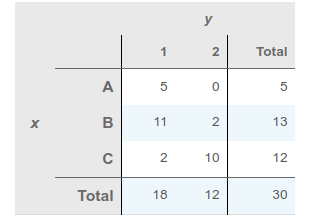
Từ bảng chéo với cột tỷ lệ phần trăm chéo của Luckett, chúng ta thấy rằng phán quyết của ông đã được giữ nguyên với tỷ lệ 91% các trường hợp tại toàn án Common Pleas và 85% tại Municipal Court. Từ bảng chéo và cột tỷ lệ phần trăm đánh giá cho Kendall, thấy rằng bản án của ông được giữ nguyên với 90% tại toà án Common Pleas và 80% tại tòa án Municipal. So sánh tỷ lệ phần trăm cột cho 2 thẩm phán, chúng ta thấy rằng Thẩm phán Luckett thể hiện kết quả tốt hơn so với Thẩm phán Kendall trong cả 2 tòa án. Kết quả này trái ngược với kết luận trước đó với Thẩm phán Kendall có kết quả tốt hơn. Ví dụ này minh họa cho nghịch lý Simpson.

Bảng chéo ban đầu được thu thập bằng cách tập hợp các dữ liệu trong bảng chéo riêng cho 2 toà án. Lưu ý rằng đối với cả 2 thẩm phán tỷ lệ kháng cáo dẫn đến đảo ngược tại tòa án Municiapl cao hơn nhiều so với tòa án Common Pleas. Như vậy, bảng chéo ban đầu, chúng ta thấy rằng loại tòa án là một biến ẩn mà không thể bỏ qua khi đánh giá các hồ sơ của 2 thẩm phấn.

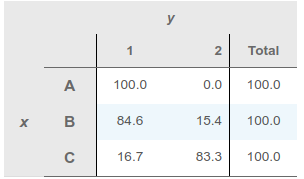
**Bài tập**

26)

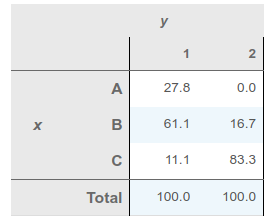
a) Bảng chéo với ***x*** là dòng, ***y*** là cột



b) Phần trăm hàng



c) Phần trăm cột



d) Mối quan hệ giữa ***x*** và ***y***:

**A** luôn có giá trị y = 1

**B** thường chủ yếu có y = 1

**C** thường chủ yếu có y = 2