*Statistics Corner* *มุมสถิติ*

# ผลที่เกิดจากการแบ่งกลุ่มตัวแปรต่อเนื่อง

**อรุณ จิรวัฒน์กุล**
*ภาควิชาชีวสถิติและประชากรศาสตร์ คณะสาธารณสุขศาสตร์ มหาวิทยาลัยขอนแก่น*

ในการวิเคราะห์ข้อมูลที่เป็นตัวแปรต่อเนื่อง เช่น อายุ ความดันโลหิต ฯลฯ พบว่ามีการวิเคราะห์ในสองรูปแบบคือ นำค่าข้อมูลต่อเนื่องมาวิเคราะห์โดยตรง เช่น คำนวณหาค่าเฉลี่ย ส่วนเบี่ยงเบนมาตรฐาน คำนวณค่า ๙๕% ช่วงเชื่อมั่นของค่าเฉลี่ย ฯลฯ อีกแบบหนึ่งทำโดยการนำข้อมูลต่อเนื่องมาแบ่งให้เป็นกลุ่มก่อนแล้วจึงนำไปวิเคราะห์ เช่นอายุแบ่งเป็นกลุ่มอายุ แล้วนำไปหาความถี่ หรือไปหาความสัมพันธ์กับการเกิดโรคจำแนกตามกลุ่มอายุ

ทำไมจึงมีการแบ่งกลุ่มให้กับข้อมูลต่อเนื่อง โดยปรกติเหตุผลของการจัดกลุ่มให้กับข้อมูลต่อเนื่องแบ่งได้เป็น ๒ ประการคือ

๑. เพื่อช่วยให้มีความเข้าใจในการนำเสนอข้อมูล เช่น ถ้าต้องการนำเสนอเพื่อเน้นว่าปัญหาของเด็กที่เป็นไข้เลือดออกมีมากในบางกลุ่มอายุ โดยการนำเสนอว่าร้อยละ ๕๔ ของเด็กที่ป่วยด้วยโรคไข้เลือดออกมีอายุต่ำกว่า ๒ ปี จะช่วยทำให้เข้าใจมากกว่านำเสนอว่าเด็กที่ป่วยด้วยโรคไข้เลือดออกมีอายุเฉลี่ย ๕.๖ ปี และส่วนเบี่ยงเบนมาตรฐาน ๔.๙ ปี หรือในกรณีที่ต้องการนำเสนอเพื่อดูความถี่ตามช่วงอายุของกลุ่มเป้าหมายที่มารับบริการ หรือเพื่อดูลักษณะการกระจายของข้อมูล เช่น ฮีสโตแกรม ปิรามิดประชากร เป็นต้น

๒. เพื่อใช้หาปัจจัยเสี่ยงหรือความสัมพันธ์ในกรณีที่พบว่าตัวแปรต่อเนื่องที่ตัวแปรต้นไม่ได้มีความสัมพันธ์เชิงเส้นตรงกับตัวแปรผล (outcome) เช่น ความเสี่ยงของการเป็นโรคหัวใจไม่ได้เพิ่มขึ้นตามอายุที่เพิ่มขึ้นในแต่ละปี แต่จะมีความเสี่ยงเพิ่มขึ้นเป็นช่วงของกลุ่มอายุ เช่น <๔๐ ปี ๔๑-๖๐ ปี และ >๖๐ ปี นอกจากนี้การแบ่งกลุ่มอายุจะช่วยทำให้ค่า odds ratio ที่คำนวณได้สามารถแปรผลตามลักษณะทางคลินิกทำให้เข้าใจได้ง่าย

ในการกำหนดจุดตัดในการแบ่งกลุ่มตัวแปรต่อเนื่องถ้าเป็นไปได้ต้องกำหนดจากเหตุผลทางคลินิกหรือทางวิทยาศาสตร์ เช่น จุดตัดที่ได้จากเหตุผลทางคลินิกของความดันโลหิตซีสโตลิก (>๑๖๐ mmHg และ <=๑๖๐ mmHg) ที่นำไปหาความสัมพันธ์กับโรคเส้นโลหิตแตกในสมอง

ในกรณีที่ยังไม่มีข้อมูลทางคลินิกหรือทางวิทยาศาสตร์ในการกำหนดจุดตัด วิธีการทางระบาดวิทยาที่ง่ายที่สุดคือ การใช้ quartiles ในการกำหนดจุดตัด ซึ่งจะแบ่งจำนวนตัวอย่างออกเป็น ๕ กลุ่มเท่า ๆ กัน นอกจากนี้ยังมีการกำหนดจุดตัดโดยใช้วิธีการทางสถิติคำนวณหาจุดตัด optimal ที่สามารถทำนายความสัมพันธ์ได้ดีที่สุด หรือการที่นักวิจัยกำหนดจุดตัดเองหลาย ๆ แบบเพื่อหาจุดตัดที่มีความสัมพันธ์มากที่สุด

ปัญหาที่พบเมื่อมีการใช้ข้อมูลกลุ่มที่ได้จากการแบ่งกลุ่มตัวแปรต่อเนื่องมาใช้ในการวิเคราะห์ทางสถิติมีดังนี้

๑. จะทำให้ผลการวิเคราะห์สูญเสียอำนาจการทดสอบ (power) ในการระบุความต่างการทดสอบของการเปรียบเทียบ และความกระชับ (precision) ในการประมาณค่าพารามิเตอร์ของประชากร เช่น ค่าเฉลี่ย odds และ hazards ฯลฯ

๒. ในการสรุปผลด้วยสถิติอนุมานจะทำให้ความผิดพลาดชนิดที่ ๑ (error) มีค่ามากกว่าที่ตั้งไว้ (๐.๐๕)[๓]

๓. การกำหนดจุดตัดแบ่งกลุ่มเอง อาจมีผลทำให้ข้อมูลจากตัวอย่างชุดเดียวกันมีโอกาสพบทั้งความสัมพันธ์เชิงลบ และความสัมพันธ์เชิงบวก

๔. ความไม่เหมาะสมในการแปลผล เช่นในการวัดคุณภาพชีวิตด้วยการทดสอบ ๑๐๐ ข้อ โดยมีระบบการให้คะแนนข้อละ ๑ คะแนน ผลการวัดที่ได้จากแต่ละบุคคลจะมีค่าตั้งแต่ ๐ (ศูนย์) ถึง ๑๐๐ คะแนน ซึ่งจะทำให้ข้อมูลที่ได้ต้องใช้วิธีการวิเคราะห์แบบข้อมูลต่อเนื่อง ถ้านักวิจัยแบ่งคะแนนที่ได้เป็นกลุ่ม ๆ โดยใช้ค่าเฉลี่ยและส่วนเบี่ยงเบนมาตรฐานออกเป็น ๕ กลุ่ม เช่นค่าต่ำสุด ๒๐ สูงสุด ๔๖ ค่าเฉลี่ย ๓๒.๗ และส่วนเบี่ยงเบนมาตรฐาน ๔.๕ คะแนน จะได้ช่วงคะแนนในแต่ละกลุ่มดังนี้ ๒๐-๒๓.๗ ๒๓.๘-๒๘.๒ ๒๘.๓-๓๗.๒ ๓๗.๓-๔๑.๗ และ ๔๑.๘-๔๖ โดยกำหนดการแปรผลทั้ง ๕ กลุ่มเป็น น้อยมาก น้อย ปานกลาง ดี และดีที่สุด การกำหนดกลุ่มโดยใช้ค่าสถิติแบ่งตามลักษณะการกระจายของข้อมูลของตัวอย่าง ไม่ได้สะท้อนเกณฑ์การวัดคุณภาพชีวิตที่มาจากเหตุผลทางวิชาการ เช่นต้องได้คะแนนตั้งแต่กี่คะแนนจึงจะถือว่ามีคุณภาพชีวิตดี ถ้าจุดตัดพิจารณาจากข้อมูลของตัวอย่างจะเห็นได้ว่ากลุ่มดีที่สุดได้คะแนนไม่ถึงครึ่งหนึ่งของคะแนนรวม ซึ่งอาจทำให้การแปลผลไม่ตรงความจริง

๕. ในการกำหนดจุดตัดโดยไม่มีเหตุผลทางคลินิก หรือทางวิทยาศาสตร์สนับสนุน มีโอกาสเกิดอคติในการแบ่งกลุ่ม เช่น แบ่งระดับการขาดเหล็กเป็นสองกลุ่ม คือ กลุ่มขาดเหล็กกับกลุ่มปรกติ ถ้าอาหารเสริมเหล็กมีผลเฉพาะกลุ่มที่ขาดเหล็กมาก (severe) เมื่อนำมาคำนวณค่าความสัมพันธ์จะทำให้ได้ข้อสรุปที่ไม่ตรงกับความจริง

**สรุป** การแบ่งกลุ่มข้อมูลตัวแปรต่อเนื่องช่วยให้การนำเสนอข้อมูลสามารถสื่อลักษณะข้อมูลได้ชัดเจนและเข้าใจได้ง่าย จุดตัดที่ใช้แบ่งกลุ่มจะต้องกำหนดจากเหตุผลทางคลินิกหรือทางวิทยาศาสตร์ ในกรณีที่จำเป็นต้องดูความสัมพันธ์ของตัวแปรที่ไม่ได้มีความสัมพันธ์เชิงเส้นตรง ถ้าต้องการอนุมานเพื่อสรุปผลความสัมพันธ์ในประชากร ควรใช้เทคนิคทางสถิติในการแปลงข้อมูลให้มีความสัมพันธ์เชิงเส้นตรง หรือการใช้ Non-linear modeling ในการวิเคราะห์

## บรรณานุกรม


๑. Altman DG. "Categorizing continuous variables," In Armitage P, Colton T, editors. Encyclopedia of biostatistics. Chichester: John Wiley; 1998; 563-7.

๒. Altman DG, Lausen B, Sauerbrei W, Schumacher M. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. Nat Cancer Inst 1994; 86:829-35.

๓. Holl N, Sauerbrei W, Schumacher M. Confidence intervals for the effect of a prognostic factor after selection of an "optimal" cutpoint. Stat Med 2004; 23:1701-13.

๔. D'Brien SM. Cutpoint selection for categorizing a continuous predictor. Biometrics 2004; 60:504-9.