วารสารวิชาการสาธารณสุข
ปีที่ 24 ฉบับที่ 2 มีนาคม - เมษายน 2558

Journal of Health Science
Vol. 24 No. 2, March - April 2015

มุมสถิติ *Statistics Corner*

การวิเคราะห์ข้อมูลที่มีการแจกแจงไม่ปกติ

อรุณ จิรวัฒน์กุล วท.บ. (อาชีวอนามัย), วท.ม. (ชีวสถิติ), M.Sc. (Clinical Epidemiology)

การทดสอบสมมติฐานและการประมาณค่าตัวแปรต่อเนื่องที่วัดด้วย interval หรือ ratio scale โดยปกติจะใช้สถิติแบบ parametric เพราะวิเคราะห์จากข้อมูลจริง และมีวิธีการทางสถิติครอบคลุมการวิเคราะห์เพื่อตอบคำถามงานวิจัยได้ทุกประเภทคำถาม แต่การใช้สถิติแบบ parametric มีข้อตกลงเบื้องต้นว่า ข้อมูลที่นำมาทดสอบสมมติฐานหรือประมาณค่า ต้องมาจากประชากรที่มีการแจกแจงปกติ

การวิเคราะห์ข้อมูลงานวิจัยส่วนใหญ่ ถ้าพบว่าประชากรมีการแจกแจงไม่ปกติ นักวิจัยมักใช้สถิติแบบ non-parametric ทดสอบสมมุติฐาน โดยไม่คำนึงถึงทางเลือกอื่น และข้อจำกัดในการแปลผล

การทดสอบสมมุติฐานแบบ non-parametric มีข้อจำกัดที่สำคัญคือ

1) การคำนวณค่าสถิติใช้ลำดับที่ไม่ใช่ค่าข้อมูลจริง ทำให้ผลสรุปที่ได้ขาดรายละเอียดความต่างที่แท้จริงของข้อมูล

2) ข้อสรุปที่ได้จากการทดสอบสมมุติฐานคือความต่างของค่ามัธยฐาน (median) ไม่ใช่ค่าเฉลี่ย (mean) การที่จะสรุปว่าเมื่อค่ามัธยฐานไม่ต่างกัน ค่าเฉลี่ยจะไม่ต่างกันด้วย ทำได้ในกรณีที่ประชากรมีการแจกแจงปกติที่ค่ามัธยฐานเท่ากับค่าเฉลี่ย

3) ค่า 95% ช่วงเชื่อมั่นเป็นของลำดับที่ไม่ใช่ช่วงข้อมูลจริง จึงไม่มีประโยชน์ในการนำสรุปผลงานวิจัย

4) การทดสอบสมมุติฐานค่าตัวแปรต่อเนื่องที่สามารถแปลงค่าให้มีการแจกแจงปกติสถิติ non-parametric มี power of test ต่ำกว่า parametric เมื่อใช้ทดสอบสมมุติฐานกรณีความแตกต่างระหว่างกลุ่มต่างกันไม่มาก ถ้าผลการทดสอบสมมุติฐานด้วยสถิติ parametric พบว่าต่างอย่างมีนัยสำคัญ แต่ผลการทดสอบด้วยสถิติ non-parametric อาจพบว่าไม่ต่าง

การวิเคราะห์ข้อมูลเมื่อผลการทดสอบสมมุติฐานลักษณะการแจกแจง (normalization test) พบว่าประชากรมีการแจกแจงไม่ปกติ นักวิจัยควรดำเนินการดังนี้

1) ทำการแปลงข้อมูล[1] (data transformations) โดยใช้วิธีการทางคณิตศาสตร์ แปลงข้อมูลจากหน่วยวัดหนึ่งไปเป็นอีกหน่วยวัดหนึ่ง ซึ่งค่าที่แปลงไปแล้ว สามารถแปลงกลับมาได้ค่าเท่าเดิม เช่น รายได้ (บาท) มีการแจกแจงเบ้ขวา แปลงเป็น Log(รายได้) จะมีการแจกแจงปกติ ซึ่งสามารถใช้วิธีวิเคราะห์แบบ parametric ทดสอบสมมติฐาน และประมาณค่าข้อมูล Log(รายได้)

ผลสรุปที่ได้จากข้อมูลแปลงจะเป็นผลต่างของค่าเฉลี่ย ผลการทดสอบสมมุติฐานของข้อมูลแปลงจะสามารถใช้สรุปผลการวิจัยได้เหมือนกับการวิเคราะห์ด้วยข้อมูลของหน่วยวัดเดิมทุกประการ และค่า 95% ช่วงเชื่อมั่นที่คำนวณจากข้อมูลแปลงสามารถแปลงกลับเป็นช่วงของหน่วยวัดเดิม จึงสามารถนำไปใช้สรุปผลวิจัยได้

2) การทดสอบสมมติฐาน หรือประมาณค่าทำบนการแจกแจงของ sampling distribution ในกรณีที่ขนาดตัวอย่างใหญ่พอ sampling distribution จะมีการ

แจกแจงปกติ (central limit theorem[1]) ในกรณีข้อมูลตัวแปรต่อเนื่องวัดด้วย interval หรือ ratio scale ที่ตัวอย่างมีขนาดใหญ่กว่า 30 จึงสามารถใช้การทดสอบแบบ parametric ในการสรุปผล

การดำเนินการในลักษณะนี้ที่พบบ่อยและทำกันมาก คือ ข้อมูลกลุ่มที่มีสองค่า (dichotomous data) ประชากรมีการแจกแจงแบบทวินาม (binomial) ถ้าตัวอย่างมีขนาดใหญ่ $np>5$ และ $n(1-p)>5$, Sampling distribution จะมีการแจกแจงปกติ จึงใช้สถิติ Z ซึ่งเป็น parametric ทดสอบสมมติฐาน และประมาณค่า ถ้าตัวอย่างมีขนาดเล็กจะใช้ Fisher's exact test (ข้อมูลกลุ่มนำมาจัดอันดับไม่ได้ จึงคำนวณ P value ด้วย exact probability โดยตรง)

3) กรณีการแจกแจงไม่ปกติ ที่ไม่สามารถแปลงข้อมูลให้มีการแจกแจงแบบปกติได้ และตัวอย่างมีขนาดเล็ก นักวิจัยจึงควรใช้ non-parametric ในการทดสอบสมมุติฐาน

สรุป

การทดสอบสมมติฐาน และการประมาณค่าข้อมูลจากประชากรที่มีการแจกแจงไม่ปกติ นักวิจัยควรพิจารณาทางเลือกต่าง ๆ ในการวิเคราะห์ก่อนที่จะใช้ non-parametric และควรใช้ สถิติ non-parametric เพื่อทดสอบสมมุติฐานเท่านั้น

เอกสารอ้างอิง

1. อรุณ จิรวัฒน์กุล. สถิติทางวิทยาศาสตร์สุขภาพเพื่อการวิจัยที่ใช้ในงานวิจัยทางวิทยาศาสตร์สุขภาพ. พิมพ์ครั้งที่ 1. กรุงเทพมหานคร: วิทยพัฒน์; 2552.