วารสารวิชาการสาธารณสุข
ปีที่ 30 ฉบับเพิ่มเติม 2 กรกฎาคม - สิงหาคม 2564

Journal of Health Science
Vol. 30 Supplement 2, July - August 2021

มุมสถิติ

Statistics Corner

ข้อควรระวังในการแปลผล p value (2)

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

บทความข้อควรระวังในการแปลผล p value (1) ในวารสารวิชาการสาธารณสุข ปีที่ 30 ฉบับที่ 4 ได้นำเสนอการแปลผล p value ไม่ถูกต้อง 3 กรณี ในบทความนี้จะนำเสนอการแปลผล p value ที่ไม่ถูกต้องเพิ่มต่ออีก 3 กรณี

กรณีที่ 4 เป็นการสรุปผล p value ที่ไม่ถูกต้องที่พบบ่อยมาก คือการทดสอบสมมุติฐานเมื่อพบว่าแตกต่างอย่างมีนัยสำคัญทางสถิติแล้ว สรุปว่าความต่างนั้นมีประโยชน์ในการนำไปใช้งาน หรือมีความสำคัญทางคลินิก (clinical importance) เป็นการสรุปที่ไม่ถูกต้องเพราะ p value บอกได้ว่ามีความต่างอย่างมีนัยสำคัญหรือไม่เท่านั้น ไม่สามารถระบุขนาดความต่างได้ จึงไม่สามารถบอกได้ว่าขนาดความต่างที่พบมีประโยชน์มากน้อยเท่าไร

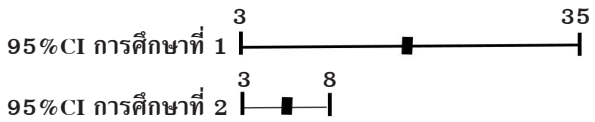
ตัวอย่างการสอนให้ผู้ป่วยเรื่องการล้างไตทางหน้าท้อง โดยมีคะแนนเต็มรวมทุกส่วน 100 คะแนน ในการทดลองเพื่อเปรียบเทียบวิธีการสอนแบบกลุ่ม (3-5 คน) และแบบบุคคลโดยใช้ตัวอย่างในการศึกษากลุ่มละ 50 คน ผลการศึกษาพบว่า Mean (SD) ของความรู้ที่สอนแบบบุคคล = 87.4 (3.2) แบบกลุ่ม = 85.6 (3.8) ผลการทดสอบสมมุติฐานได้ p value = 0.01 แสดงว่าวิธีการสอนแบบบุคคลได้ผลดีกว่าแบบกลุ่มอย่างมีนัยสำคัญ ผู้วิจัยเสนอให้ใช้วิธีสอนแบบบุคคลเป็นมาตฐานในการสอนผู้ป่วยกลุ่มนี้ กรณีนี้เป็นการนำความแตกต่างทางสถิติที่พบไปสรุปว่ามีประโยชน์ในการนำไปใช้งาน ซึ่งไม่ถูกต้องเพราะเมื่อพิจารณาความต่างจากค่าช่วงเชื่อมั่นพบว่าสองกลุ่มต่างกัน 1.8 คะแนน CI [0.4, 3.2] การสอนสองวิธีได้ผลต่างกัน 1.8 คะแนนจากคะแนนเต็ม 100 ถือว่าแตกต่างกันน้อยมาก เมื่อดูคะแนนเฉลี่ย พบว่าทั้งสองกลุ่มมีคะแนนความรู้สูงเกิน 85% แสดงว่าทั้งสองกลุ่มมีความรู้ดี ซึ่งความรู้ดีต่างกัน 1.8 คะแนนไม่น่าจะทำให้การดูแลตัวเองได้ต่างกัน และการสอนแบบกลุ่มยังช่วยลดภาระงานของเจ้าหน้าที่ได้อีกด้วย ดังนั้นควรสรุปว่า “ผลการศึกษาพบว่าคะแนนความรู้เฉลี่ยทั้งสองกลุ่มมีความรู้ดีเกินร้อยละ 85 ถึงแม้ทั้งสองกลุ่มมีความรู้ต่างกันอย่างมีนัยสำคัญ แต่ขนาดความต่าง 1.8 คะแนนไม่น่าจะทำให้การดูแลตัวเองได้ต่างกันจึงควรให้ใช้วิธีการสอนแบบกลุ่มตามเดิม”

เนื่องจากขนาดตัวอย่างเป็นปัจจัยสำคัญที่ทำให้ความแตกต่างขนาดเล็กให้ผลการทดสอบที่ให้ p value ต่ำ ดังนั้นการศึกษาที่มีตัวอย่างขนาดใหญ่ เช่น การสำรวจขนาดใหญ่ การศึกษาที่ใช้ข้อมูลจากทะเบียนโรค เมื่อพบความต่างอย่างมีนัยสำคัญควรเพิ่มความระมัดระวังในการแปลผล

การพิจารณาขนาดความต่างที่มีประโยชน์ จะพิจารณาจากขนาดความต่างที่พบเพียงอย่างเดียวไม่ได้ นักวิจัยจำเป็นต้องมีความรู้เกี่ยวกับเนื้อหาในเรื่องนั้น มีรู้เรื่องกระบวนการและเวลาที่ใช้ในการเปลี่ยนแปลงผลลัพธ์ ถ้าผู้วิจัยเป็นผู้ปฏิบัติงานที่เกี่ยวข้องกับเรื่องดังกล่าวยิ่งจะช่วยให้กำหนดขนาดความต่างที่เป็นประโยชน์ได้สอดคล้องกับความจริง

กรณีที่ 5 ผลการทดสอบสมมุติฐานที่ได้ p value เท่ากันแล้วสรุปว่าการทดสอบทั้งระบุขนาดความต่างได้เท่ากัน เป็นการสรุปที่ไม่ถูกต้อง

ตัวอย่าง การศึกษาที่ 1 เปรียบเทียบ treatment A กับกลุ่มควบคุม การศึกษาที่ 2 treatment B กับกลุ่มควบคุม ผลการทดสอบสมมุติฐานพบว่าทั้งสองการศึกษามี p value = 0.38 เท่ากัน ผู้วิจัยจะสรุป ว่า treatment A และ B ทั้งสองให้ผลต่างจากกลุ่มควบคุมเท่ากัน ข้อสรุปดังกล่าวไม่ถูกต้อง เพราะ p value ที่คำนวณได้นอกจากจะขึ้นอยู่กับขนาดความต่างแล้ว ยังขึ้นอยู่กับขนาดตัวอย่างในการศึกษาด้วย ถ้าต้องการดูขนาดความต่างจะต้องพิจารณาจากค่าช่วงเชื่อมั่น ซึ่งเมื่อคำนวณ 95% ช่วงเชื่อมั่นของ treatment effect พบว่า การศึกษาที่ 1 มีค่าอยู่ระหว่าง 3% ถึง 35% ส่วนการศึกษาที่ 2 มีค่าอยู่ระหว่า 3% ถึง 8% ดังภาพข้างล่าง



จากภาพ treatment effect ของการศึกษาทั้งสองไม่เท่ากัน โดยการศึกษาที่ 2 ช่วงเชื่อมั่นกระชับแต่ขนาดความต่างค่อนข้างน้อย ส่วนการศึกษาที่ 1 ขนาดความต่างค่อนข้างมาก แต่ตัวอย่างมีขนาดเล็กทำให้ช่วงเชื่อมั่นกว้าง

ดังนั้นการทดสอบสมมุติฐานที่มี p value เท่ากันอาจมีขนาดความต่าง ๆ ระหว่างกลุ่มไม่เท่ากัน ในกรณีผลสรุปพบว่าต่างกันอย่างมีนัยสำคัญ p value เท่ากันบอกได้ว่า การสรุปทั้งสองการทดสอบมีโอกาสสรุปผิดจากความบังเอิญเท่ากัน

กรณีที่ 6 คือการที่นักวิจัยทำการทดสอบสมมุติฐานที่เป็นไปได้ทั้งหมดเพื่อหาตัวแปรที่พบความแตกต่างอย่างมีนัยสำคัญ (p value<0.05) เป็นกรณีที่พบบ่อยเช่นกัน เป็นการวิเคราะห์ที่เรียกว่า data fishing หรือ data dredging ที่มีการทดสอบสมมุติฐานจำนวนมาก ซึ่งมีผลทำให้การสรุปผลการทดสอบในภาพรวมมีค่า type I error มากกว่าค่า α ที่กำหนด ดังนั้นถ้าต้องให้ผลสรุปรวมของการทดสอบทั้งหมด มีความผิดพลาดเท่ากับค่า α ที่ตั้งไว้ นักวิจัยต้องลดขนาดค่า α ของการทดสอบแต่ละครั้ง เช่นเดียวกันกับที่ใช้ในการทำ multiple comparison

สำหรับงานวิจัยแบบ exploratory ที่ต้องการดูว่าตัวแปรใดมีแนวโนมที่จะมีความสัมพันธ์กับตัวแปรผล จะมีการทดสอบสมมุติฐานที่เป็นไปได้ทั้งหมดเช่นเดียวกัน แต่สรุปผลจาก exploratory research มุ่งที่จะประเมินความสัมพันธ์ของตัวแปรแต่ละตัวเพื่อนำไปพิสูจน์ซ้ำ จึงไม่นับเป็น data fishing

สรุป

การแปลผล p value ต้องระลึกอยู่เสมอว่า p value บอกได้เพียงว่าต่างกันอย่างมีนัยสำคัญหรือไม่เท่านั้น ไม่สามารถระบุขนาดความต่าง การระบุขนาดความต่างที่พบมีประโยชน์ ต้องพิจารณาจากช่วงเชื่อมั่น

การทดสอบที่มี p value เท่ากันไม่สามารถบอกได้ว่ามีขนาดความต่างเท่ากัน ในกรณีที่พบว่าต่างอย่างมีนัยสำคัญ p value ที่เท่ากันสรุปได้ว่ามีโอกาสสรุปผิดจากความบังเอิญเท่ากัน

ในการวิเคราะห์ข้อมูลควรทำการทดสอบสมมุติฐานเฉพาะที่จำเป็น เพราะการทดสอบสมมุติฐานหลายครั้งที่เกินความจำเป็นทำให้ การสรุปผลวิจัยในภาพรวมมี type I error มากกว่าค่า α ที่กำหนด