

การแปลผล P value ไม่ถูกต้อง

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

การนำเสนอผลการทดสอบสมมติฐานทางสถิติสามารถนำเสนอด้วย p value และช่วงเชื่อมั่น การนำเสนอด้วย p value จะให้ข้อมูลเพื่อตัดสินใจว่าผลการทดสอบ (เปรียบเทียบ) ต่างกันอย่างมีหรือไม่มีนัยสำคัญ ส่วนช่วงเชื่อมั่น นอกจากจะบอกว่าต่างกันอย่างมีหรือไม่มีนัยสำคัญแล้ว ยังสามารถระบุขนาดและทิศทางของความต่างได้ด้วย มีการแปลความหมาย p value ผิดในหลายลักษณะ จึงมีข้อแนะนำในแนวปฏิบัติการเขียนรายงานวิจัย เช่น CONSORT⁽¹⁾ ให้รายงานผลการทดสอบสมมติฐานด้วยช่วงเชื่อมั่น แต่การนำเสนอด้วย p value ยังคงมีใช้อยู่เหมือนเดิมไม่ได้ลดลง⁽²⁾

ในบทความวิจัยภาษาไทย การทดสอบสมมติฐานนำเสนอด้วยช่วงเชื่อมั่นน้อยมาก เกือบทั้งหมดนำเสนอด้วย p value และยังพบว่ามีการแปลผลที่ไม่ถูกต้องอยู่เสมอ

ก่อนอธิบายลักษณะการแปลผล p value ที่ไม่ถูกต้องควรทำความเข้าใจที่มาของ p value และการใช้ในการตัดสินใจผลการทดสอบสมมติฐาน p value คำนวณโดยมีข้อตกลงเบื้องต้นว่า สมมติฐานไม่ต่าง (null hypothesis) เป็นจริงในประชากร เช่น ในการทดสอบความต่างระหว่างค่าเฉลี่ยของประชากร A กับ B ค่า p value จะคำนวณโดยมีข้อตกลงเบื้องต้นว่าสมมติฐานไม่ต่างเป็นจริง หมายถึงค่าเฉลี่ยของประชากรทั้งสองไม่ต่างกัน

ถ้า $\bar{X}_A = 37.2$ และ $\bar{X}_B = 38.6$ เป็นค่าเฉลี่ยของตัวอย่างที่สุ่มได้จากประชากร A และ B ซึ่งพบว่า ค่าเฉลี่ยของตัวอย่างทั้งสองต่างกัน ในการทดสอบสมมติฐานเพื่อสรุปว่าค่าเฉลี่ยของประชากรทั้งสองต่างกันหรือไม่ ทำโดยคำนวณค่าสถิติ t แล้วเปลี่ยนเป็นค่า p value ถ้าได้ $p=0.12$ ซึ่งมีค่ามากกว่าค่า α (0.05) จะยอมรับสมมติฐาน สรุป

ว่าค่าเฉลี่ยของประชากร A และ B ต่างกันอย่างไม่มีนัยสำคัญ ความต่างที่พบอาจมาจากความบังเอิญ (chance) ของการสุ่มตัวอย่าง

ในกรณีที่คำนวณได้ $p=0.04$ ซึ่งมีค่าน้อยกว่าค่า α (0.05) จะปฏิเสธสมมติฐานสรุปว่าค่าเฉลี่ยของประชากร A และ B ต่างกันอย่างมีนัยสำคัญ ความต่างที่พบไม่ได้เกิดจากความบังเอิญ

การนำค่า p value ไปเทียบกับเกณฑ์ตัดสินใจ (α) จะได้ผลสรุปว่าความต่างที่พบมาจากความบังเอิญหรือไม่ การยอมรับ หรือปฏิเสธ สมมติฐาน จะนำไปแปลผลได้เพียงว่าค่าเฉลี่ยของประชากรต่างกันอย่างมีนัยสำคัญหรือไม่เท่านั้น ไม่สามารถระบุขนาดและทิศทางของความต่าง

ในบทความวิจัยภาษาไทยการแปลผล p value ไม่ถูกต้องที่พบบ่อย มี 4 ลักษณะคือ (1) การสรุปความต่างทางสถิติที่พบว่ามีประโยชน์ในการใช้งาน หรือมีความสำคัญทางคลินิก (clinical importance) (2) p value มีค่าน้อย ยิ่งต่างกันมาก (3) ผลการทดสอบพบว่าต่างกันอย่างไม่มีนัยสำคัญแล้วไปสรุปว่าไม่มีความต่าง และ (4) ความไม่เหมาะสมในการนำเสนอค่า p value

ความไม่เหมาะสมลักษณะแรกพบได้บ่อยมาก รายงานวิจัยส่วนใหญ่เมื่อนำค่า p value ไปเทียบกับค่า α แล้วพบว่าต่างอย่างมีนัยสำคัญ จะสรุปว่าความต่างที่พบว่ามีประโยชน์ต่อการใช้งานโดยไม่พิจารณาข้อมูลอื่นประกอบ เช่น วิธีการสอนใหม่ทำให้มีความรู้มากขึ้นจาก 61 เป็น 64 คะแนน และการทดสอบพบว่าต่างกันอย่างมีนัยสำคัญ นักวิจัยสรุปผลว่าวิธีการสอนใหม่ทำให้มีความรู้มากกว่าวิธีที่ใช้อยู่เดิม ควรนำวิธีการสอนใหม่นี้ไปใช้แทนวิธีการสอนเดิม

ถ้าพิจารณาจากขนาดความต่างของความรู้ที่เพิ่มขึ้น 3 คะแนนจากคะแนนเต็ม 100 คะแนน อาจไม่มีผลต่อการเปลี่ยนพฤติกรรม จึงไม่น่าจะมีประโยชน์พอที่จะนำไปใช้งาน การแปลผล p value จะสรุปได้ว่าความต่างที่พบไม่ได้มาจากความบังเอิญ การพิจารณาประโยชน์ในการใช้งานต้องพิจารณาจากขนาดความต่างและระดับคะแนนประกอบด้วย เช่น แบบวัดความรู้ แบบวัดทัศนคติ หรือแบบวัดพฤติกรรม ควรพิจารณาร่วมกับคะแนนเต็ม (เพื่อระดับคะแนน) และขนาดความต่าง เช่น ความรู้มีคะแนนเต็ม 100 คะแนน วิธีการสอนใหม่ทำให้มีความรู้มากขึ้นจาก 81 เป็น 90 ถึงแม้พบว่าต่างกันมาก แต่วิธีการสอนที่ใช้อยู่เดิมได้ผลดีมากอยู่แล้ว การเปลี่ยนวิธีการสอนใหม่อาจจะได้ประโยชน์เพิ่มขึ้นไม่มากนัก ทั้งยังต้องลงทุนในการฝึกอบรม และผู้ใช้ที่ต้องปรับตัวกับการเปลี่ยนวิธีสอนใหม่

P value ที่คำนวณได้อาจมีค่า $p < 0.05$, $p < 0.01$ หรือ $p < 0.001$ ความไม่ถูกต้องในลักษณะที่ 2 เกิดจากการสรุปว่าค่า p value มีค่าน้อย ยิ่งต่างกันมาก เช่น ผลการทดสอบสมมุติฐานเพื่อเปรียบเทียบค่าเฉลี่ยได้ $p < 0.001$ นักวิจัยสรุปว่าค่าเฉลี่ยระหว่างกลุ่มต่างกันมาก การสรุปแบบนี้ไม่ถูกต้องเพราะขนาดของ p value จะแปรตามขนาดความต่างและขนาดตัวอย่าง ขนาดความต่างเท่าเดิม ถ้าขนาดตัวอย่างเพิ่มขึ้น p value จะมีค่าเล็กลง ดังนั้นจึงไม่สามารถสรุปขนาดความต่างจากค่า p value การแปลผลที่ถูกต้องคือ p value มีค่าน้อย ยิ่งช่วยเพิ่มความมั่นใจในการสรุปว่า ความต่างที่พบไม่ได้เกิดจากความบังเอิญ

ความไม่ถูกต้องในลักษณะที่ 3 เกิดเมื่อผลการทดสอบพบว่าต่างกันอย่างไม่มีความสำคัญ แล้วนักวิจัยสรุปว่าค่าเฉลี่ยของประชากรทั้งสองเท่ากัน ในกรณีพบว่าต่างกันอย่างไม่มีความสำคัญ นักวิจัยจะสรุปว่า ค่าเฉลี่ยของประชากรเท่ากันไม่ได้ เพราะค่า p value ยังขึ้นอยู่กับขนาดตัวอย่างในกรณีความต่างเท่าเดิม ถ้าเพิ่มขนาดตัวอย่างให้ใหญ่ขึ้น ค่า p value จะมีขนาดเล็กลงจนพบความต่างที่มีนัยสำคัญได้ การสรุปที่ถูกต้องควรพิจารณาขนาดความต่าง และขนาดตัวอย่างประกอบด้วย เพื่อดูว่า

ขนาดความต่างที่พบมีขนาดเล็ก (ค่าเฉลี่ยมีค่าใกล้เคียงกัน) หรือขนาดตัวอย่างที่ใช้ในการศึกษาไม่ใหญ่พอที่จะระบุความต่าง ซึ่งพิจารณาได้จากช่วงเชื่อมั่น

ความไม่เหมาะสมลักษณะที่ 4 เป็นเรื่องการนำเสนอค่า p value ความไม่เหมาะสมที่พบคือการนำเสนอค่า p value เป็น 0.000 หรือ เป็นทศนิยมเกิน 3 ตำแหน่ง

P value จะมีค่าเป็นศูนย์ไม่ได้ โปรแกรมวิเคราะห์บางโปรแกรมรายงานค่า p value เป็น 0.000 ที่เป็นเช่นนี้เกิดจากการปัดค่าทศนิยมให้เหลือ 3 ตำแหน่ง ดังนั้น ถ้านำผลนี้ไปเขียนในรายงานวิจัย ควรนำเสนอเป็น $p < 0.001$ p value ควรนำเสนอเป็นเลขทศนิยมไม่เกิน 3 ตำแหน่ง เพราะเพียงพอสำหรับการเปรียบเทียบกับค่า α จะนำเสนอด้วยค่าที่คำนวณได้โดยตรง เช่น $p = 0.03$ หรือ $p = 0.017$ หรือนำเสนอในรูปของค่ามากกว่า หรือน้อยกว่าค่า α ที่ใช้เป็นเกณฑ์ตัดสินก็ได้ เช่น $p < 0.05$, $p < 0.01$ หรือ $p < 0.001$ การจะเลือกวิธีเขียนแบบใดนักวิจัยควรดูคำแนะนำรูปแบบการเขียนจากวารสารที่จะตีพิมพ์

สรุป

การนำเสนอผลการทดสอบสมมุติฐานทางสถิติ ด้วย p value เมื่อพบว่ามีขนาดความต่างอย่างมีนัยสำคัญ การระบุประโยชน์ในการใช้งาน ควรพิจารณาขนาดความต่างร่วมด้วย ในกรณีที่พบว่าต่างอย่างไม่มีความสำคัญ ควรพิจารณาขนาดความต่าง และขนาดตัวอย่างประกอบกันในการสรุปผล วิธีการนำเสนอผลการทดสอบสมมุติฐานควรนำเสนอทั้ง p value และช่วงเชื่อมั่น ในกรณีนำเสนอด้วย p value อย่างเดียวให้ระมัดระวังในการแปลผล

เอกสารอ้างอิง

1. Moher D, Hopewell S, Schulz KF, Montori V, Gotzsche PC, Devereaux PJ, et al. CONSORT 2010 Explanation and Elaboration: updated guidelines for reporting parallel group randomised trials. *BMJ* 2010;340:c869.
2. Chavalarias D, Wallach JD, Li AHT, Ioannidis JPA. Evolution of reporting p values in the biomedical literature, 1990–2015. *JAMA* 2016;315:1141–8.